Assigning Priorities (or not) in Service Systems with Nonlinear Waiting Costs

Huiyin Ouyang
HKU Business School, The University of Hong Kong, Hong Kong
Nilay Tanık Argon, Serhan Ziya
Department of Statistics and Operations Research, UNC Chapel Hill, Chapel Hill, NC 27599
oyhy@hku.hk, nilay@unc.edu, ziya@unc.edu

For a queueing system with multiple customer types differing in service-time distributions and waiting costs, it is well known that the $c\mu$-rule is optimal if costs for waiting are incurred linearly with time. In this paper, we seek to identify policies that minimize the long-run average cost under nonlinear waiting cost functions within the set of *fixed priority policies* that only use the type identities of customers independently of the system state. For a single-server queueing system with Poisson arrivals and two or more customer types, we first show that some form of the $c\mu$-rule holds with the caveat that the indices are complex, depending on the arrival rate, higher moments of service time, and proportions of customer types. Under quadratic cost functions, we provide a set of conditions that determine whether to give priority to one type over the other or not to give priority but serve them according to first-come-first-served (FCFS). These conditions lead to useful insights into when strict (and fixed) priority policies should be preferred over FCFS and when they should be avoided. For example, we find that when traffic is heavy, service times are highly variable, and the customer types are not heterogenous, then prioritizing one type over the other (especially a proportionally dominant type) would be worse than not assigning any priority. By means of a numerical study, we generate further insights into more specific conditions under which fixed priority policies can be considered as an alternative to FCFS.

History: This paper was first submitted in January 2018, first revision submitted in August 2019, and second revision submitted in August 2020.

1. Introduction

Many service systems prioritize their customers based on customers' characteristics such as expected service time and value to the system in addition to their arrival times to the system. For example, patients arriving at the emergency department of a hospital are first triaged, i.e., assigned a criticality level, and prioritized based on their triage category and arrival order. Another example is call centers, where customers who have premier membership status are given priority for order of service. A natural framework for analyzing such systems has been through modeling them as queueing systems and over the last sixty years numerous articles have been published on how customers in a queueing system should be prioritized.

Despite significant progress, however, this literature still has important gaps from both academic and practical points of view largely due to the assumptions imposed on the waiting costs for analytical tractability. Specifically, an overwhelming proportion of prior work assumes that the cost of waiting for a customer is a linear function of the customer's waiting time, an assumption that is not likely to hold for many systems. For example, the optimality of the well-known $c\mu$-rule has been established under a variety of conditions but all under the restriction that waiting costs are linear (see Cox and Smith (1961) for the article that started this literature and see Section 2 for more on the $c\mu$-rule). On the other hand, the work that considered the possibility of nonlinear waiting costs imposed some other restrictions on the system such as the requirement that the system operate under heavy traffic and the waiting cost function be convex. More importantly, the policies proposed (e.g., the generalized $c\mu$-rule by Van Mieghem (1995)) are somewhat sophisticated requiring the system to keep track of the queue-waiting time of each customer and to have complete knowledge of the waiting cost function, which may pose a challenge in practice.

While prioritization is prevalent in practice, in many cases, the policies in place are not based on careful statistical estimation of the waiting cost functions but mostly based on some rough analysis of limited data, and the service providers' past experience and beliefs about who needs the

service more urgently or whose long wait would be more detrimental for the system. For example, prioritization of patients in emergency departments or in the aftermath of mass-casualty events is very common in practice yet the precise nature of the effect of passage of time on the patient survival, which can be seen as waiting cost, is not well understood (see Jenkins et al. (2008), Sacco et al. (2005) and the discussion on survival probability functions in Sun et al. (2017)). Similarly, in other settings like healthcare clinics and call centers, there is very limited work on the estimation of waiting costs. Nevertheless, this does not stop providers from implementing prioritization policies that they believe to be improving system performance. They also usually stick with simple policies like classifying customers into a few groups and prioritizing one group over the other without taking into account state of customers. In this paper, we call such a policy *a fixed priority policy* because the priority order assigned to each customer type does not change with time.

Given the fact that waiting cost functions are not known precisely, choosing a fixed priority policy as opposed to dynamically prioritizing customers based on exact cost information (if one needs to be chosen) is reasonable. But the question remains as to whether fixed prioritization makes sense in the first place. The theory supports prioritization among classes when waiting costs are linear functions of time but what if the waiting costs are not linear? When is there at least some justification for taking the risk of using prioritization between classes and thereby possibly alienating customers rather than using a standard first-come-first-served (FCFS) policy, which is at least largely perceived to be fair? A provider who uses prioritization without knowing the precise form of the waiting cost functions is in fact implicitly assuming a certain relationship between the waiting cost functions for different classes. But what are these implicit assumptions? One of the two main goals of this article is to provide some answers to these questions, which we do by comparing the performance of applying FCFS across different types with those of assigning fixed priorities under cost functions that are not necessarily linear.

The second goal of this article is to provide some managerial insights into the type of conditions that would favor a particular strict priority policy or FCFS over other fixed priority policies. (In this paper, a *strict* priority policy is a fixed priority policy under which there is at least one type that is prioritized over others under all circumstances.) While service providers might find it difficult to estimate the waiting cost functions precisely, they might have a good sense of the general structure of the function (convex, concave, quadratic, etc.). Thus, it would be useful to know, assuming that the cost functions have a particular structure (but not knowing the functions precisely), whether any one of the policies would stand out by being the best choice under a larger or more realistic set of cost parameter values than the others and whether the policy that stands out depends on system conditions such as traffic intensity. For example, if a linear cost model appears to be appropriate for most customers but a quadratic cost function for one particular class, would any one of the policies stand out as more likely to be better than the others? Would the answer depend on the traffic intensity on the system? How about the service time variability?

In the pursuit of the goals stated above, we analyzed an M/G/1 queueing system with two or more types of customers and each type being characterized by a service time distribution and a waiting cost function, where the waiting cost function for at least one type is nonlinear. The performance measure of interest is the long-run average cost, and hence, priority policies that provide a smaller performance measure are better. Following a review of the relevant literature in Section 2, we provide more details of our stylized model in Section 3.

Our theoretical analysis starts with a set of conditions that determine the order between three fixed priority policies that differ only in the priority orders of two types of customers under a general cost structure; see Section 4. Although these conditions may not be any simpler than directly comparing long-run average costs under competing fixed priority policies, they demonstrate that the comparison follows some form of the famous $c\mu$-rule. In Section 5, we continue our theoretical analysis by taking a closer look at the case with quadratic waiting cost functions, which generates several interesting and useful insights. For example, we find that the choice between priority policies depends on the traffic intensity and the proportion of each type in the population unlike in the

linear-cost case. We also provide results on how the decision to prioritize or not to prioritize changes with the composition of the population and traffic intensity. To further strengthen these insights, we study the case with partial information on the waiting cost functions in Section 6 and present results of a numerical study in Section 7. We provide the most important conclusions from this study in Section 8. The proofs of our analytical results, some supplemental material, and tables of notation are provided in the Appendix.

2. Literature review

Queueing systems where certain classes of customers have priority over others are called *priority queues*. The study of priority queues dates back to Cobham (1954) who considered a single-server Markovian queueing system (M/M/1) where customers belong to multiple priority classes and the service is non-preemptive. For such a system, Cobham (1954) derived expressions for the long-run average waiting times in the queue for each priority class. This seminal work was followed by Miller (1960) and Jaiswal (1968), who advanced the analysis of priority queues further, e.g., by providing Laplace-Stieltjes transforms of the waiting time distributions for M/G/1 priority queues and considering other priority mechanisms such as preemptive prioritization. Others also considered probabilistic priority policies, where priorities are assigned randomly among different customer classes; e.g., Katayama and Takahashi (1992) and Jiang et al. (2002) provided approximations for the delay performance under such policies.

When the waiting time of customers is penalized linearly with time, Cox and Smith (1961) established the optimality of the well-known $c\mu$-rule, which minimizes the long-run average waiting cost in an M/G/1 queue with multiple priority classes. According to the $c\mu$-rule, customers with larger $c_i\mu_i$ index are assigned higher priority, where c_i is the waiting cost per unit time and μ_i is the service rate for type i customers. Following this seminal paper, the optimality of the $c\mu$-type policies has been studied under various settings by Kakalik and Little (1971), Klimov (1974, 1979), Harrison (1975), Pinedo (1983), Nain (1989), Argon and Ziya (2009), Budhiraja et al. (2014) among others, all under the assumption of linear cost functions. We also refer readers who are interested in more general conditions for the optimality of $c\mu$-type policies to research on achievable regions for optimal control of queueing systems, e.g., Shanthikumar and Yao (1992) and Bertsimas (1995).

While this is not the first paper to consider nonlinear waiting costs in queueing systems, it would be fair to say that the literature on the topic is scarce. Within this literature, Haji and Newell (1971) showed that when waiting cost functions are increasing and convex, the optimal policy will always serve customers of the same type according to the FCFS discipline. Later, Van Mieghem (1995) proved that when waiting costs are convex in time, a generalized version of the $c\mu$-rule is asymptotically optimal under heavy traffic, which was followed by a proof by Mandelbaum and Stolyar (2004) that extended the heavy-traffic optimality of the generalized $c\mu$-rule to more general settings. The generalized $c\mu$-rule is a dynamic policy that gives priority to the customer who has the largest $C_i'(t)\mu_i$ value in the system at every service completion epoch, where $C_i(t)$ is the cost of a type i customer with a queue-waiting time of t units and $C_i'(t)$ is its first-order derivative. Hence, to implement the generalized $c\mu$-rule, one needs to keep track of the waiting times of all customers in the system and know the cost functions precisely.

Other relevant work that study the optimal scheduling problem in priority queueing systems under convex cost structures include Ansell et al. (2003), Glazebrook et al. (2003), and Bispo (2013). Assuming that the holding cost is a function of the number of customers in the system, these papers developed state-dependent (dynamic) heuristic policies for single-server queueing systems as an alternative to the simpler generalized $c\mu$-rule. Gurvich and Whitt (2009) considered a multi-server multi-class service system with convex delay costs that are functions of the queue length. They introduced a queue-and-idleness-ratio policy and showed that this proposed policy would reduce to the $c\mu$-rule under linear holding costs and to the generalized $c\mu$-rule under strictly convex costs and other regularity conditions. Finally, Ata and Tongarlak (2013) and Larranaga et al. (2015) studied the dynamic control of multi-class queueing systems with abandonments and proposed state-dependent heuristic policies that would work under possibly nonlinear waiting costs.

3. Model description

Consider a single-server queueing system with K types of customers, where K is an integer and $2 \le K < \infty$. Customers arrive to the system according to a Poisson process with rate $\lambda > 0$, and each arriving customer belongs to type $i \in \{1, 2, \ldots, K\}$ with probability $p_i > 0$, where $\sum_{i=1}^{K} p_i = 1$, independently of the arrival process. Service times for type i customers are independent and identically distributed (i.i.d.) with rate $\mu_i > 0$ and nth moment $\tau_i^{(n)} > 0$ for $n \ge 2$. We define $\mu \equiv 1/\sum_{i=1}^{K}(p_i/\mu_i)$, $\tau^{(n)} \equiv \sum_{i=1}^{K} p_i \tau_i^{(n)}$, $\rho_i \equiv \lambda p_i/\mu_i$, and $\rho \equiv \lambda/\mu$, which we call the traffic intensity, and we assume that $\rho < 1$ for stability. Each type i customer incurs a waiting cost $C_i(t)$ when its waiting time in the queue is $t \ge 0$. We assume that $C_i(t)$ is first-order differentiable and non-decreasing in t for fixed i. (See Appendix A.1 for tables of notation used throughout this paper.)

For such a queueing system, we consider a set of policies Π that only includes non-idling and non-preemptive queueing policies that assign a *fixed (deterministic) priority order* to each type of customers. More specifically, Π consists of policies under which customers are ranked according to at most K priority orders and any policy $\pi \in \Pi$ satisfies the following properties: Let K^π be the number of distinct priority orders under policy π. Without loss of generality, let $\{1, 2, \ldots, K^\pi\}$ denote the set of priority orders under policy π and assume that a smaller priority order represents a higher priority for service. (Different customer types may have the same priority order, and hence, $1 \le K^\pi \le K$.) Priority orders are *fixed* in the sense that they cannot be modified once the system starts operating. A customer from a type with priority order $k > 1$ cannot be taken into service when there exists a customer in the system that has a priority order smaller than k. Policies in Π are also non-idling and non-preemptive in the sense that the server does not idle as long as there is a customer in the system and that service of a customer who has been taken into service has to be completed without any preemption before the server moves on to serving another customer.

For any policy $\pi \in \Pi$, define the long-run average cost as

$$C_\pi \equiv \lim_{t \to \infty} \frac{\sum_{i=1}^{K} \sum_{k=1}^{n_i(t)} C_i(V_{i,k}^{\pi,x_0})}{t} \tag{1}$$

(whenever the limit exists), where $n_i(t)$ is the number of type i arrivals by time t and $V_{i,k}^{\pi,x_0}$ is the waiting time of the kth arriving type i customer under policy π and initial state x_0. Our objective is to compare policies in Π in terms of their long-run average waiting costs where smaller the long-run average cost better the policy. Let W_i^π denote the steady-state waiting time of a type i customer under policy π. We show in the Appendix that the limit in (1) exists and satisfies

$$C_\pi = \lambda \sum_{i=1}^{K} p_i E\Big[C_i(W_i^\pi)\Big] \tag{2}$$

if Assumption 1 holds.

Assumption 1. For $\pi \in \Pi$ and all $i = 1, 2, \ldots, K$, $E\Big[\big|C_i(W_i^\pi)\big|\Big] < \infty$.

In this paper, we focus on policies in the subset $\Pi^F \subset \Pi$, where customers with the same priority order are served according to FCFS. These policies are of special interest due to their common use in practice. Furthermore, if $C_i(\cdot)$ is a convex function (in the non-strict sense) for all $i = 1, 2, \ldots, K$, then it is sufficient to compare only policies in Π^F to find an optimal policy within Π. More precisely, Proposition 1 in Haji and Newell (1971) implies that if $C_i(t)$ is convex for all i, then, for any policy $\pi \in \Pi \setminus \Pi^F$, there exists a policy $\pi^F \in \Pi^F$ such that $C_{\pi^F} \le C_\pi$. (Proposition 1 in Haji and Newell (1971) assumes two types but can be easily extended to more than two types.)

We conclude this section by noting that Assumption 1 holds under some reasonable conditions for the cost functions and service time distributions as long as the queueing system is stable. As an example, see Proposition 1 below for polynomial cost functions.

PROPOSITION 1. *Suppose that for type $i \in \{1,2,\ldots,K\}$ customers, $C_i(t) = \sum_{\ell=1}^{J_i} c_i^{(\ell)} t^\ell$, where $J_i < \infty$ is the degree of the polynomial function $C_i(t)$ and $c_i^{(\ell)}$ are some real numbers such that $C_i'(t) \geq 0$ for all $t \geq 0$. If $\rho < 1$ and the first $J_i + 1$ moments of service times for all customers are finite, then $E[(W_i^\gamma)^\ell] < \infty$ for $\ell = 1,2,\ldots,J_i$ and $\gamma \in \Pi^F$, and hence, $E[|C_i(W_i^\gamma)|] < \infty$.*

The condition on the moments of service times in Proposition 1 holds for many common distributions such as exponential, gamma, Weibull, and lognormal.

4. Comparison under general cost functions: $c\mu$-rule appears again

We start the analysis by comparing policies in Π^F under very general conditions for the waiting cost functions (i.e., Assumption 1). To understand the complexity of the problem, first consider the case with two types, under which Π^F has three policies: one that prioritizes type 1 customers, another that prioritizes type 2 customers, and a third one that prioritizes neither type and follows FCFS for all customers. When there are three or more types, the number of ways of prioritizing and/or "pooling" different types of customers will increase dramatically. (In this paper, pooling means grouping two or more customer types and assigning the same priority order to the group.) Specifically, there are 13 choices with three types of customers: prioritize each individual type (six choices), pool two types and assign priorities between the pooled group and the single type (six choices), or pool all three. Note that the number of policies in Π^F corresponds to the ordered Bell number in number theory, which can be approximated by $K!/\left(2(\log 2)^{K+1}\right)$, see, e.g., Gross (1962). This shows that it becomes much more difficult to compare all policies in Π^F as K increases.

Although we are not able to compare all policies in Π^F analytically (except for $K = 2$), as we show in the remainder of this paper, we can compare policies that are similar in that they share the same priority order assignment for all types except for two. Such comparisons not only provide a means to eliminate suboptimal policies, but more importantly, also generate insights into how priorities should be assigned between two types, especially when the priority orders of other types are pre-determined.

To present our results, we start by picking a policy $\pi \in \Pi^F$. Without loss of generality, we assume that type i customers have priority order i under policy π, for $i = 1,2,\ldots,K$. For fixed $k \in \{1,2,\ldots,K-1\}$, we will compare π with two of its variants. We first define policy π_{k+1} to be a priority policy where the priority orders of all types are the same as those under policy π except that the priority orders of type k and type $k+1$ customers are switched. We also define policy $\bar{\pi}_k$ to be a priority policy where the priority orders of all types of customers are the same as those under policy π except that type k and type $k+1$ customers are pooled and served according to FCFS. Thus, policies π_{k+1} and $\bar{\pi}_k$ that are derived from policy π treat all types the same way except for types k and $k+1$: policy π_{k+1} prioritizes type $k+1$ customers over type k and $\bar{\pi}_k$ does not differentiate between types k and $k+1$ customers in terms of priority. For notational convenience, we also define policy π_k, which is identical to policy π. Our next result provides necessary and sufficient conditions for the comparison of π_k, π_{k+1}, and $\bar{\pi}_k$ under general cost structures.

PROPOSITION 2. *Suppose that Assumption 1 holds under policies π_k, π_{k+1}, and $\bar{\pi}_k$ for fixed $k \in \{1,2,\ldots,K-1\}$, and let*

$$\delta_i^{\gamma_1,\gamma_2} \equiv \frac{E\left[C_i(W_i^{\gamma_1})\right] - E\left[C_i(W_i^{\gamma_2})\right]}{E\left[W_i^{\gamma_1}\right] - E\left[W_i^{\gamma_2}\right]}, \tag{3}$$

for $i \in \{k, k+1\}$, $\gamma_1, \gamma_2 \in \{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, and $\gamma_1 \neq \gamma_2$. Then, we have:
(a) $C_{\pi_k} \leq C_{\bar{\pi}_k}$ if and only if $\delta_k^{\pi_k,\bar{\pi}_k}\mu_k \geq \delta_{k+1}^{\pi_k,\bar{\pi}_k}\mu_{k+1}$;
(b) $C_{\pi_{k+1}} \leq C_{\bar{\pi}_k}$ if and only if $\delta_{k+1}^{\pi_{k+1},\bar{\pi}_k}\mu_{k+1} \geq \delta_k^{\pi_{k+1},\bar{\pi}_k}\mu_k$; and
(c) $C_{\pi_k} \leq C_{\pi_{k+1}}$ if and only if $\delta_k^{\pi_k,\pi_{k+1}}\mu_k \geq \delta_{k+1}^{\pi_k,\pi_{k+1}}\mu_{k+1}$.

Proposition 2 is not difficult to prove as can be seen in the Appendix but it has an elegant interpretation: *the ordering among long-run average costs under π_k, π_{k+1}, and $\bar{\pi}_k$ follows a generalized version of the famous cμ-rule.* To see this more clearly, note that $\delta_i^{\gamma_1,\gamma_2}$ defined in (3) can also be expressed as $\delta_i^{\gamma_1,\gamma_2} = E[C_i'(U_i^{\gamma_1,\gamma_2})]$, where $U_i^{\gamma_1,\gamma_2}$ is a random variable that is defined by a probabilistic analogue of the mean value theorem (see Remark A.1 in the Appendix). In particular, $U_i^{\gamma_1,\gamma_2}$ can be considered as the "mean value" of steady-state waiting times $W_i^{\gamma_1}$ and $W_i^{\gamma_2}$. Thus, $\delta_i^{\gamma_1,\gamma_2}$ can be interpreted as the expected marginal change in cost for type i customers when the policy switches from γ_1 to γ_2, or vice versa. Then, parts (a) and (b) of Proposition 2 say that prioritizing type i customers over type j customers, for $i,j \in \{k,k+1\}$ and $i \neq j$, is better than pooling them if and only if switching from policy $\bar{\pi}_k$ to π_i results in a larger expected marginal decrease in cost per unit service time for type i customers than the expected marginal increase in cost per unit service time for type j customers. Similarly, part (c) of Proposition 2 says that prioritizing type k over type $k+1$ is better than the opposite if and only if switching from policy π_{k+1} to π_k results in a larger expected marginal decrease in cost per unit service time for type k customers than the expected marginal increase in cost per unit service time for type $k+1$.

As long as precise expressions for the cost functions for the two types under comparison are known, it is not difficult to numerically determine $\delta_i^{\gamma_1,\gamma_2}$, and thus, to identify the best policy within $\{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, which is Π^F itself for $K=2$. Indeed, as we show in Corollary A.1 in the Appendix, we only need to compute $\delta_i^{\gamma,\bar{\pi}_k}$ for $i \in \{k,k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$ to find the best policy among π_k, π_{k+1}, and $\bar{\pi}_k$ using Proposition 2. Hence, in the rest of this paper, we simplify the notation by letting $\delta_i^{\gamma} \equiv \delta_i^{\gamma,\bar{\pi}_k}$ for $i \in \{k,k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$. Furthermore, under certain assumptions on the structure of the waiting cost functions, it is possible to obtain closed-form expressions for δ_i^{γ} as we demonstrate for quadratic functions in Section 5. Finally, note that the computation of δ_i^{γ} for $i \in \{k,k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$ does not require knowledge of cost functions of other types but only those of the two types of customers we compare, which is practically appealing.

5. Comparison under quadratic cost functions

Polynomial waiting cost functions – especially quadratic costs – have been widely used in the study of queueing systems with nonlinear waiting costs; see, e.g., Ata and Tongarlak (2013) and Parlar and Sharafali (2014). These functions have been popular not only because they are suitable for analysis but also because they fit well to the perceived cost of waiting in several applications. For example, in a recent empirical study, Ding et al. (2019) find that the marginal waiting cost of critical patients (from the decision makers' perspective) at four large Canadian emergency departments can be approximated by a piece-wise linear increasing function. With this motivation, we focus on polynomial cost functions with a degree of at most two in the rest of this paper to derive more insights into our main research question: when to assign priorities if waiting costs are nonlinear?

To apply Proposition 2 to the quadratic-cost case, we need to define

$$M_i^{\gamma} \equiv \frac{E\left[\left(W^{\bar{\pi}_k}\right)^2\right] - E\left[\left(W_i^{\gamma}\right)^2\right]}{E\left[W^{\bar{\pi}_k}\right] - E\left[W_i^{\gamma}\right]}, \tag{4}$$

for $i \in \{k,k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$. (We drop the subscript from $W_i^{\bar{\pi}_k}$ for $i \in \{k,k+1\}$ because $W_k^{\bar{\pi}_k}$ and $W_{k+1}^{\bar{\pi}_k}$ are identical.) In words, M_i^{γ} represents the change in second moments of steady-state waiting times of type i customers per change in their mean steady-state waiting times by switching from priority policy γ to $\bar{\pi}_k$. As can be seen in Proposition 3 below, we need M_i^{γ} to characterize the best policy among π_k, π_{k+1}, and $\bar{\pi}_k$ under quadratic costs. Closed form expressions of M_i^{γ} are given in Equations (EC.8) and (EC.9) in the Appendix and show that M_i^{γ} is a function of the arrival rate, proportions of customer types, and first three moments of service times.

Proposition 3. *Suppose that Assumption 1 holds, and for fixed $k \in \{1,2,\ldots,K-1\}$ and $i \in \{k,k+1\}$, cost functions satisfy $C_i(t) = c_i^{(2)}t^2 + c_i^{(1)}t$, where $t \geq 0$ and $c_i^{(1)}, c_i^{(2)} \geq 0$. Then, $\delta_i^{\pi_i} \leq \delta_i^{\pi_j}$*

for $i,j \in \{k,k+1\}$ and $j \neq i$, where $\delta_i^\gamma = c_i^{(2)} M_i^\gamma + c_i^{(1)}$ for $\gamma \in \{\pi_k, \pi_{k+1}\}$, and the inequality holds strictly if and only if $c_i^{(2)} > 0$. Furthermore, the best policy among π_k, π_{k+1}, and $\bar{\pi}_k$ is characterized as follows:
(a) if $\delta_k^{\pi_k}\mu_k > \delta_{k+1}^{\pi_k}\mu_{k+1}$, then π_k is the best (and π_{k+1} is the worst);
(b) if $\delta_{k+1}^{\pi_{k+1}}\mu_{k+1} > \delta_k^{\pi_{k+1}}\mu_k$, then π_{k+1} is the best (and π_k is the worst); and
(c) otherwise, i.e., if $\delta_k^{\pi_k}\mu_k \leq \delta_{k+1}^{\pi_k}\mu_{k+1}$ and $\delta_{k+1}^{\pi_{k+1}}\mu_{k+1} \leq \delta_k^{\pi_{k+1}}\mu_k$, $\bar{\pi}_k$ is the best.

Proposition 3 is a generalization of the classical $c\mu$-rule to the quadratic-cost setting. (One can recover the $c\mu$-rule by setting $c_i^{(2)} = 0$ for all $i = 1, 2, \ldots, K$ and applying Proposition 3 multiple times.) Possibly the most important difference from the classical $c\mu$-rule is that in the quadratic-cost case prioritizing one type or the other may be worse than not prioritizing. (By the first statement of Proposition 3 on the strict order between $\delta_i^{\pi_i}$ and $\delta_i^{\pi_j}$ and part (c), we know that there is a non-empty region where $\bar{\pi}_k$ is the best.) In particular, when there are only two types of customers, this result suggests that FCFS can be better than prioritizing either type under quadratic cost functions while FCFS is suboptimal under linear costs. The reason is that when costs are linear, unlike in the non-linear case, each additional unit of waiting adds the same amount to the total cost regardless of how long the wait has been. Hence, in the case of non-linear costs, priority among customers should depend not only on their types but also on how long they have waited. (This has been noted by others including Van Mieghem (1995).) Therefore, a fixed deterministic policy that prioritizes one type over the other can lead to excessive waits for the non-priority type whereas this would not be the case in FCFS due to the randomized order of arrivals of different types. Proposition 3 also states that serving both types k and $k+1$ according to FCFS, i.e., not giving priority between types k and $k+1$, is never the worst policy. This can be best explained by the strict order between $\delta_i^{\pi_i}$ and $\delta_i^{\pi_j}$ for $i,j \in \{k,k+1\}$ and $i \neq j$, because it implies that the expected marginal decrease in cost for type i by switching from $\bar{\pi}_k$ to π_i is (strictly) less than the expected marginal increase in cost for the same type by switching from $\bar{\pi}_k$ to π_j. In other words, when costs are quadratic, *the harm caused by giving lower priority to a type is more than the benefit gained by prioritizing it in terms of its expected marginal cost.*

To gain further insights, we next study the case where $c_k^{(1)}\mu_k = c_{k+1}^{(1)}\mu_{k+1}$, e.g., when $c_k^{(1)} = c_{k+1}^{(1)} = 0$. (Argon et al. (2009) and Ata and Tongarlak (2013) are examples of work that studied similar cost structures.)

ASSUMPTION 2. For $k \in \{1, 2, \ldots, K-1\}$ and $i \in \{k, k+1\}$, we have $C_i(t) = c_i^{(2)}t^2 + c_i^{(1)}t$, where $c_i^{(1)} \geq 0, c_i^{(2)} > 0$, and $c_k^{(1)}\mu_k = c_{k+1}^{(1)}\mu_{k+1}$.

COROLLARY 1. *Under Assumptions 1 and 2, the best policy among π_k, π_{k+1}, and $\bar{\pi}_k$ is characterized as follows: π_{k+1} is the best (and π_k is the worst) if $c_k^{(2)}\mu_k/(c_{k+1}^{(2)}\mu_{k+1}) < R^{\pi_{k+1}}$; π_k is the best (and π_{k+1} is the worst) if $c_k^{(2)}\mu_k/(c_{k+1}^{(2)}\mu_{k+1}) > R^{\pi_k}$; and $\bar{\pi}_k$ is the best if $R^{\pi_{k+1}} \leq c_k^{(2)}\mu_k/(c_{k+1}^{(2)}\mu_{k+1}) \leq R^{\pi_k}$, where $R^\gamma \equiv M_{k+1}^\gamma/M_k^\gamma$ for $\gamma \in \{\pi_k, \pi_{k+1}\}$ and $R^{\pi_{k+1}} < R^{\pi_k}$. Furthermore, $R^{\pi_{k+1}} < 1 < R^{\pi_k}$ if types k and $k+1$ are identical in terms of the first two moments of their service times.*

Corollary 1 completely characterizes the best policy among π_k, π_{k+1}, and $\bar{\pi}_k$ for quadratic cost functions with $c_k^{(1)}\mu_k = c_{k+1}^{(1)}\mu_{k+1}$. In particular, it states that π_k is the best if $c_k^{(2)}\mu_k$ is sufficiently larger than $c_{k+1}^{(2)}\mu_{k+1}$, π_{k+1} is the best if $c_k^{(2)}\mu_k$ is sufficiently smaller than $c_{k+1}^{(2)}\mu_{k+1}$, and $\bar{\pi}_k$ is the best if the values of $c_k^{(2)}\mu_k$ and $c_{k+1}^{(2)}\mu_{k+1}$ are not significantly different. Note particularly the non-empty optimality region for the pooled policy $\bar{\pi}_k$, i.e., where $c_k^{(2)}\mu_k/(c_{k+1}^{(2)}\mu_{k+1}) \in [R^{\pi_{k+1}}, R^{\pi_k}]$. Using the work conservation law (see, e.g., Kleinrock (1965)), R^γ can be interpreted as the ratio of changes in the second moment of steady-state waiting times for type $k+1$ customers versus type k customers when switching from priority policy γ to $\bar{\pi}_k$, adjusted by their respective traffic intensity. Corollary 1 also implies that under quadratic cost functions if all customers have the same first and second moments of service times, then $R^{\pi_{k+1}} < 1 < R^{\pi_k}$ for all $k = 1, 2, \ldots, K-1$, and hence, policies that give priority to customer types with smaller values of $c_i^{(2)}$ should not be considered.

5.1. Effects of system parameters on conditions for prioritization

In this section, we focus on the case with $K = 2$ and investigate how the intervals for $c_k^{(2)}\mu_k/(c_{k+1}^{(2)}\mu_{k+1})$ given in Corollary 1, over which one policy is better than the others, change with system parameters like the arrival rate. Such an analysis leads to insights into the question of when to prioritize and is especially useful if the service provider has reason to believe that quadratic functions would accurately capture the waiting costs but cannot determine $c_k^{(2)}$ and $c_{k+1}^{(2)}$ precisely.

More specifically, we study how the prioritization decision for the case with two types of customers changes with the traffic intensity ρ, proportion of type 1 customers p_1, and the service time distributions under quadratic cost functions $C_i(t) = c_i^{(2)}t^2$ for $c_i^{(2)} > 0$ and $i \in \{1,2\}$. First, note that when there are two types of customers in the system, π_i is the priority policy that prioritizes type $i \in \{1,2\}$ customers, $\bar{\pi}_1$ corresponds to FCFS, and $\Pi^F = \{\pi_1, \pi_2, \text{FCFS}\}$. Then, Corollary 1 provides a complete characterization of the best policy in Π^F: π_1 is the best if $c_1^{(2)}\mu_1/(c_2^{(2)}\mu_2) > R^{\pi_1}$, π_2 is the best if $c_1^{(2)}\mu_1/(c_2^{(2)}\mu_2) < R^{\pi_2}$, and FCFS is the best otherwise. We start by providing some numerical examples to visually depict these optimality regions and how they change with ρ (or equivalently with λ) and p_1.

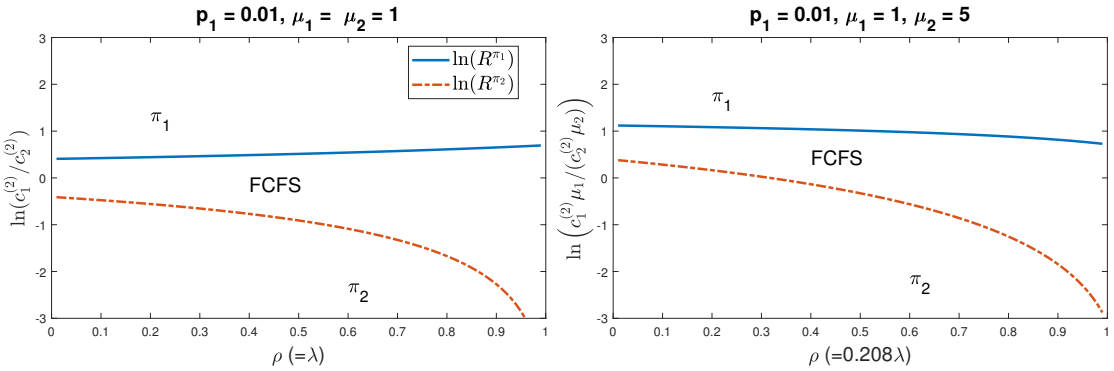

Figure 1 **Regions of optimality for FCFS, π_1, and π_2 as a function of ρ (or equivalently λ) under exponential service times and quadratic waiting costs with $C_i(t) = c_i^{(2)}t^2$ for $t \geq 0$ and $i \in \{1,2\}$.**

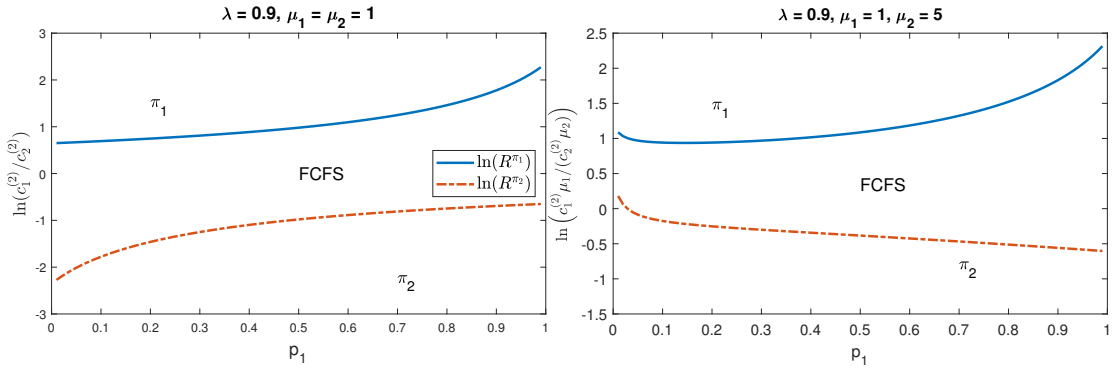

Figure 2 **Regions of optimality for FCFS, π_1, and π_2 as a function of p_1 under exponential service times and quadratic waiting costs with $C_i(t) = c_i^{(2)}t^2$ for $t \geq 0$ and $i \in \{1,2\}$.**

Figures 1 and 2 show how the optimal policy shifts from π_2 to FCFS first and then to π_1 as $c_1^{(2)}\mu_1/(c_2^{(2)}\mu_2)$ increases in agreement with Corollary 1. (Figures 1 and 2 present the most representative plots from a more detailed numerical study.) Furthermore, it appears from Figure 1 that the region where FCFS is the best policy enlarges as λ increases and that both R^{π_1} and R^{π_2}

monotonically change with λ but they are not necessarily increasing or decreasing in all cases. (By the symmetry between the two types and the right-most plot in Figure 1, we know that there is also a case where both R^{π_1} and R^{π_2} increase with ρ.) We also notice from Figure 2 that R^{π_1} and R^{π_2} do not always change monotonically in p_1 except when all service times are i.i.d. In Propositions 4 and 5, we prove some of these observations on the monotonicity of R^{π_1} and R^{π_2}, and also provide their limits in heavy traffic.

PROPOSITION 4. *(a)* R^{π_1} *increases in* λ *if* $\tau_2^{(2)}\mu_2 \geq \tau_1^{(2)}\mu_1(1-2\rho_1)/(2-2\rho_1)$.
(b) R^{π_2} *decreases in* λ *if* $\tau_1^{(2)}\mu_1 \geq \tau_2^{(2)}\mu_2(1-2\rho_2)/(2-2\rho_2)$.
(c) $R^{\pi_1} \to \dfrac{p_1\mu_1^{-1}+2p_2\mu_2^{-1}}{p_2\mu_2^{-1}}$ *and* $R^{\pi_2} \to \dfrac{p_1\mu_1^{-1}}{2p_1\mu_1^{-1}+p_2\mu_2^{-1}}$ *as* $\lambda \to \mu = (p_1/\mu_1 + p_2/\mu_2)^{-1}$.

PROPOSITION 5. *When the first three moments of service times are identical for all customers,* R^{π_1} *and* R^{π_2} *both increase in* p_1 *(and hence decrease in* p_2*).*

Propositions 4 and 5 provide several useful insights into when prioritization should be considered over FCFS and when not. In the following, we provide an itemized list of these insights, where each item first states the insight followed by how it is derived from Proposition 4 or 5. Later in Section 7.1, we observe through a numerical study that most of these insights could be generalized to systems with more than two customer types.

- *If* $\tau_i^{(2)}\mu_i$ *values for the two customer types are relatively close, specifically, one is not more than twice as large as the other, then the region where FCFS is the best gets larger as the traffic intensity grows while the optimality regions for the priority policies get smaller:* From parts (a) and (b) of Proposition 4, if $1/2 < \tau_1^{(2)}\mu_1/(\tau_2^{(2)}\mu_2) < 2$, then R^{π_1} increases and R^{π_2} decreases with λ.
- *The region where prioritizing the proportionally dominant type is the best shrinks as the arrival rate increases and the traffic intensity of that type surpasses 1/2*: If $\rho_1 \geq 1/2$ [$\rho_2 \geq 1/2$], then the condition in part (a) [(b)] of Proposition 4 is automatically satisfied, and hence, R^{π_1} increases [R^{π_2} decreases] with λ. This can be also observed in Figure 1.
- *Under heavy traffic, the boundaries of optimality regions depend only on customer mix in the population and the first moment of service times*: This is immediate from Proposition 4 (c) and is consistent with heavy traffic analysis of other queueing systems in the literature with nonlinear penalties for waiting. For example, see Van Mieghem (1995), Ata and Tongarlak (2013), Ata and Peng (2018), where order of policies depends on first moment measures (such as service rates and/or abandonment rates) and not on higher moments under heavy traffic.
- *When the first three moments of service times are identical for the two types, as the proportion of one type increases, the optimality region for giving priority to that type shrinks while the optimality region for prioritizing the other type enlarges:* This directly follows from Proposition 5 and can be also seen from the left-most plot in Figure 2.
- *Under heavy traffic, if one type significantly dominates the other in terms of proportion of population, then giving priority can only be justified for the type with the smaller proportion and that justification also requires that its* $c_i^{(2)}\mu_i$ *is at least twice as large as that of the other type. Otherwise, it is better to use FCFS:* By Proposition 4 (c), under heavy traffic, $R^{\pi_1} \to 2$ and $R^{\pi_2} \to 0$ as $p_1 \to 0$, whereas $R^{\pi_1} \to \infty$ and $R^{\pi_2} \to 1/2$ as $p_1 \to 1$. Hence, under heavy traffic, type i customers should not be prioritized if the proportion of this type is close to one; instead, the other type, i.e., type $3-i$, should be served first if $c_{3-i}^{(2)}\mu_{3-i} > 2c_i^{(2)}\mu_i$, and otherwise FCFS should be applied.

In this section, we also compare the values of R^{π_1} and R^{π_2} under two different service time distributions with the same means to observe effects of service time distributions (or equivalently second moments in the quadratic-cost case). For $\gamma \in \{\pi_1, \pi_2\}$, let R^{γ}_{exp} and R^{γ}_{det} denote the values of R^{γ} under exponential and deterministic service times for all customers, respectively.

PROPOSITION 6. *(a)* $R^{\pi_1}_{exp} \geq R^{\pi_1}_{det}$ *if and only if* $\mu_1 \geq \mu_2(1-\rho_1)/(2-\rho_1)$.
(b) $R^{\pi_2}_{exp} \leq R^{\pi_2}_{det}$ *if and only if* $\mu_1 \leq \mu_2(2-\rho_2)/(1-\rho_2)$.

Proposition 6 implies that if $\frac{\mu_2}{\mu_1} \in \left(\frac{1-\rho_2}{2-\rho_2}, \frac{2-\rho_1}{1-\rho_1}\right)$, then $R^{\pi_2}_{exp} \leq R^{\pi_2}_{det} < R^{\pi_1}_{det} \leq R^{\pi_1}_{exp}$, and hence, when the mean service times are not significantly different for the two types of customers, FCFS is preferable for a wider range of values of $c_1^{(2)}/c_2^{(2)}$ under exponential service times than under deterministic service times. This suggests that *when the two types are not too different in terms of mean service times, higher service time variability makes FCFS a better choice under a larger range of waiting cost scenarios.* When service times have higher variance, waiting times will also have higher variance regardless of whether FCFS or a strict fixed priority policy is in place. Nevertheless, due to the convexity of the waiting cost functions, the impact will be larger on the strict priority policies because of the longer waits experienced by at least some of the lower priority customers.

It is important to note however that if mean service times are sufficiently different between the two types, lower variability might make prioritizing the type with smaller mean service time more desirable. Specifically, Proposition 6 also says that if one type is sufficiently faster to serve in the mean sense, say, $\mu_2/\mu_1 > (2-\rho_1)/(1-\rho_1)$, then $R^{\pi_2}_{exp} \leq R^{\pi_2}_{det}$ and $R^{\pi_1}_{exp} \leq R^{\pi_1}_{det}$, which implies that under deterministic service times, π_2 (prioritizing the faster type) is preferred for a wider range of values of $c_1^{(2)}/c_2^{(2)}$, and π_1 (prioritizing the slower type) is preferred for a narrower range of values of $c_1^{(2)}/c_2^{(2)}$ than that under exponential service times.

We would like to conclude this section with a summary of insights that could be most useful to managers. Higher arrival rates favor FCFS over strict priority policies and these priority policies are justifiable only if there is a high level of heterogeneity between types (in terms of cost parameters, first two moments of service times, and proportions) under heavy traffic. Moreover, under heavy traffic, if a strict fixed priority is better than FCFS, then it must be the one that gives priority to the type with a smaller proportion of the customer population. Hence, when there are concerns about the heterogeneity among types or when service times are suspected to be highly variable, managers should be cautious about replacing FCFS with strict priority policies in heavily loaded systems with quadratic waiting costs.

5.2. Could prioritization be a fair policy?

In this section, we discuss the implications of our results on quadratic cost functions on the problem of minimizing the variance of steady-state waiting times when the mean service times for all customers are the same but the variance and higher moments are possibly different. Minimization of variance of steady-state waiting times has been of great interest in the context of fairness in queueing systems. In particular, Kingman (1962), Avi-Itzhak and Levy (2004), and references therein use variance of waiting times as a measure of fairness in a queueing system in that a policy that has a smaller variance of waiting times is regarded as a fairer policy. Kingman (1962) and Vasicek (1977) prove that FCFS minimizes the variance of waiting times among all non-idling queueing disciplines and thus is the "fairest" discipline for various queueing systems with a single class of customers. Avi-Itzhak and Levy (2004) propose a new fairness measure that computes the expected number of positions that a job is pushed ahead or backwards under a policy compared to FCFS but conclude that for G/G/c queues with c parallel servers, variance of the steady-state waiting time can be used as an appropriate measure of fairness. To the best of our knowledge, all earlier work on minimization of variance of waiting times considered customers from a single class. Here, we study the variance minimization problem for an M/G/1 queue with *multiple classes* of customers with equal service rates but possibly different service-time distributions.

For identical service rates for all customers, the steady-state mean waiting times are the same under any policy in Π, as can be verified by the work conservation law. Hence, minimizing the variance of the steady-state waiting times within Π is equivalent to minimizing its second moment, which corresponds to letting $C_i(t) = t^2$ $(t \geq 0)$ for all i in our formulation. Then, we can use Corollary 1 to prove the following result.

PROPOSITION 7. *Suppose that λ, μ_i, $\tau_i^{(2)}$, and $\tau_i^{(3)}$ are finite and $\mu_i = \mu > \lambda$ for all $i = 1, 2, \ldots, K$, and without loss of generality, $p_k \leq p_{k+1}$ for fixed $k \in \{1, 2, \ldots, K-1\}$. Let π_k^* be a policy that minimizes the variance of the steady-state waiting times within the set $\{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$.*

(a) If $\rho \geq \left(\sum_{j=1}^{k+1} p_j + \sqrt{(p_k + p_{k+1})p_k}\right)^{-1}$, *then* $\pi_k^* = \bar{\pi}_k$.

(b) If $\left(\sum_{j=1}^{k+1} p_j + \sqrt{(p_k + p_{k+1})p_{k+1}}\right)^{-1} \leq \rho < \left(\sum_{j=1}^{k+1} p_j + \sqrt{(p_k + p_{k+1})p_k}\right)^{-1}$, *then there exists a threshold* $\xi_k > \tau_{k+1}^{(2)}$ *such that*

$$\pi_k^* = \begin{cases} \bar{\pi}_k, & \text{if } \tau_k^{(2)} \leq \xi_k; \\ \pi_{k+1}, & \text{if } \tau_k^{(2)} > \xi_k. \end{cases}$$

(c) If $\rho < \left(\sum_{j=1}^{k+1} p_j + \sqrt{(p_k + p_{k+1})p_{k+1}}\right)^{-1}$, *then there exist thresholds* $\xi_k > \tau_{k+1}^{(2)}$ *and* $\tilde{\xi}_k < \tau_{k+1}^{(2)}$ *such that*

$$\pi_k^* = \begin{cases} \pi_k, & \text{if } \tau_k^{(2)} < \tilde{\xi}_k; \\ \bar{\pi}_k, & \text{if } \tilde{\xi}_k \leq \tau_k^{(2)} \leq \xi_k; \\ \pi_{k+1}, & \text{if } \tau_k^{(2)} > \xi_k. \end{cases}$$

Proposition 7 explicitly shows the effects of traffic intensity, proportions of types, and service time variances on the selection of the fairest policy. In particular, from Proposition 7 (a), we can see that if the traffic intensity is sufficiently large, then prioritizing either type over the other is worse than pooling these two types together. After a closer examination of this lower bound on ρ in part (a), we find that this condition could possibly hold for $\rho < 1$ if and only if the total proportion of the two types under consideration is sufficiently large (i.e., $p_k + p_{k+1} > \sqrt{2}\sum_{j=k+2}^{K} p_j$) and the dominant type (i.e., type $k+1$ because $p_k \leq p_{k+1}$) does not heavily dominate the other type in proportion (i.e., $1 \leq p_{k+1}/p_k < \left((p_k + p_{k+1})/\sum_{j=k+2}^{K} p_j\right)^2 - 1$); see proof of part (a) of Proposition 7 in the Appendix. (These two conditions are automatically satisfied when $K = 2$.) Proposition 7 (a) also implies that FCFS is better than any policy that groups the types into two priority classes if $\rho \geq \left(1 + \sqrt{\min(p_1, p_2, \ldots, p_K)}\right)^{-1}$, i.e., if traffic is sufficiently heavy and/or none of the types constitute too small a portion of the population.

Proposition 7 (b) states that if the traffic is not as heavy as in part (a) but also is not too light, then prioritizing the type with a smaller proportion can never minimize the variance of steady-state waiting times, and which of the remaining two policies is best depends on the service time variances of the two types under consideration. More specifically, since service rates are the same for types k and $k+1$, comparison of $\tau_k^{(2)}$ and $\tau_{k+1}^{(2)}$ is the same as the comparison of service time variances. Hence, under mediocre traffic intensity, prioritizing the proportionally dominant type is the best if the service time variance for the other type is significantly larger than that of the dominant type, or otherwise pooling the two types is the best.

Finally, Proposition 7 (c) shows that when the traffic is light, giving priority to one type over the other is preferable if and only if its service time variance is sufficiently smaller than that of the other type. However, when variances of service times for the two types are similar, then serving them according to FCFS can still be better than prioritizing either type even though the traffic is light. Note that we can obtain closed-form expressions for thresholds ξ_k and $\tilde{\xi}_k$ in Proposition 7, but in the interest of space, we provide these in the proof of Proposition 7 in the Appendix.

6. Prioritization under imperfect cost information

In Sections 4 and 5, we provided analytical comparisons of policies that prioritize certain types over others and those that pool them under the assumption that cost functions for the types under consideration are completely known. In certain situations, however, we may have reliable estimates on the exact functional form of waiting cost of one type but only have partial information on the waiting cost of other types such as a range of their marginal waiting costs. For example, if we use

regression models to estimate the cost function from data for different types of customers, we may not be able to obtain good regression models for all types but for certain types we can estimate a range for the marginal costs.

In this section, we present a result, namely, Corollary 2, that orders the long-run average costs under policies π_k, π_{k+1}, and $\bar{\pi}_k$, which are defined in Section 4 for fixed $k \in \{1, 2, \ldots, K\}$, when we have only partial cost information on one of the two types under consideration. Specifically, we assume that $C_k(t)$ is completely known, but we only know the range of values that $C'_{k+1}(t)$ could take. (Corollary 2 also holds if we switch the indices k and $k+1$).

COROLLARY 2. *Suppose that Assumption 1 holds under policies π_k, π_{k+1}, and $\bar{\pi}_k$.*
(a) If $C'_{k+1}(t) \geq \max\{\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}}\}\mu_k/\mu_{k+1}$ for all $t \geq 0$, then $C_{\pi_{k+1}} \leq C_{\bar{\pi}_k} \leq C_{\pi_k}$.
(b) If $C'_{k+1}(t) \leq \min\{\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}}\}\mu_k/\mu_{k+1}$ for all $t \geq 0$, then $C_{\pi_k} \leq C_{\bar{\pi}_k} \leq C_{\pi_{k+1}}$.
(c) If $\delta_k^{\pi_k}\mu_k/\mu_{k+1} \leq C'_{k+1}(t) \leq \delta_k^{\pi_{k+1}}\mu_k/\mu_{k+1}$ for all $t \geq 0$, then $C_{\bar{\pi}_k} \leq C_{\pi_{k+1}}$ and $C_{\bar{\pi}_k} \leq C_{\pi_k}$.

Corollary 2 provides bounds on $C'_{k+1}(t)$ for all $t \geq 0$, namely, $\delta_k^{\pi_k}\mu_k/\mu_{k+1}$ and $\delta_k^{\pi_{k+1}}\mu_k/\mu_{k+1}$, to compare policies π_k, π_{k+1}, and $\bar{\pi}_k$. Specifically, $C'_{k+1}(t)$ has to be bounded from either above or below for all $t \geq 0$ (which is true, for example, when the cost function is concave) for the conditions of the corollary to hold. We next look at two special cases to demonstrate how this result could be useful.

Quadratic cost for one type: Suppose type k customers are known to have a quadratic cost function, i.e., $C_k(t) = c_k^{(2)}t^2 + c_k^{(1)}t$ for $c_k^{(1)} \geq 0$ and $c_k^{(2)} > 0$, but we do not know the exact form of $C_{k+1}(t)$. In this case, we have $\delta_k^{\gamma} = c_k^{(2)}M_k^{\gamma} + c_k^{(1)}$, where M_k^{γ} is given by (4) for $\gamma \in \{\pi_k, \pi_{k+1}\}$ and $\delta_k^{\pi_k} < \delta_k^{\pi_{k+1}}$ (see the proof of Proposition 3). Then, Corollary 2 implies that if the waiting cost for type $k+1$ customers increases at a sufficiently large rate at all times, i.e., $C'_{k+1}(t)$ is at least $\delta_k^{\pi_{k+1}}\mu_k/\mu_{k+1}$, then type $k+1$ customers should be prioritized over type k; if the waiting cost of type $k+1$ customers increases at a small rate at all times, i.e., $C'_{k+1}(t)$ is at most $\delta_k^{\pi_k}\mu_k/\mu_{k+1}$, then type k customers should be prioritized over type $k+1$; and if the derivative of $C_{k+1}(t)$ lies between $\delta_k^{\pi_k}\mu_k/\mu_{k+1}$ and $\delta_k^{\pi_{k+1}}\mu_k/\mu_{k+1}$ at all times, then assigning the same priority to these two types is the best. Furthermore, we notice that $\delta_k^{\pi_k}$, $\delta_k^{\pi_{k+1}}$, and the difference $\delta_k^{\pi_{k+1}} - \delta_k^{\pi_k}$ all increase in λ (see (EC.8), (EC.9), and (EC.10) in the Appendix), which implies that the bounds $\delta_k^{\pi_k}\mu_k/\mu_{k+1}$ and $\delta_k^{\pi_{k+1}}\mu_k/\mu_{k+1}$, and the length of the interval $(\delta_k^{\pi_k}\mu_k/\mu_{k+1}, \delta_k^{\pi_{k+1}}\mu_k/\mu_{k+1})$ are all increasing as λ becomes larger. Indeed, both $\delta_k^{\pi_k}$ and $\delta_k^{\pi_{k+1}}$ go to infinity as λ approaches μ for $k = K-1$, which leads to an important conclusion: *if one type has a quadratic cost function and the derivatives of the cost functions of all the other types are bounded from above, then it is never the best to assign the lowest priority to the type with quadratic cost when the traffic intensity is heavy no matter what the service time and cost parameters are for the other types.*

Linear cost for at least one type: Suppose that type k customers are known to have a linear cost function, i.e., $C_k(t) = c_k t$ for $t \geq 0$ and $c_k > 0$, but we do not know the exact form of $C_{k+1}(t)$. Then, we have $\delta_k^{\pi_k} = \delta_k^{\pi_{k+1}} = c_k\mu_k$, and by Corollary 2, if $C'_{k+1}(t) \geq c_k\mu_k/\mu_{k+1}$ for all $t \geq 0$, then $C_{\pi_{k+1}} \leq C_{\bar{\pi}_k} \leq C_{\pi_k}$; and if $C'_{k+1}(t) \leq c_k\mu_k/\mu_{k+1}$ for all $t \geq 0$, then $C_{\pi_k} \leq C_{\bar{\pi}_k} \leq C_{\pi_{k+1}}$. This means that even if we do not know the exact waiting cost function for type $k+1$ customers, if we know that their marginal waiting cost at any amount of wait is greater than [less than] $c_k\mu_k/\mu_{k+1}$, then prioritizing type $k+1$ [type k] customers is better than prioritizing the other type or pooling these two types. This result also leads to another practical finding for systems that have linear cost functions for all types but one. More specifically, suppose that $C_j(t) = c_j t$ for $t \geq 0$, $c_j > 0$, and $j = 2, 3, \ldots, K$. If $C'_1(t) \geq \max_{j=2,\ldots,K}\{c_j\mu_j\}/\mu_1$ for all $t \geq 0$, then Corollary 2 implies that type 1 should receive the highest priority and all other types should follow the $c\mu$-rule. Similarly, if $C'_1(t) \leq \min_{j=2,\ldots,K}\{c_j\mu_j\}/\mu_1$ for all $t \geq 0$, then type 1 should receive the lowest priority while all other types follow the $c\mu$-rule. To demonstrate further, consider convex-concave (or S-shaped) cost functions that are commonly discussed in service operations literature. If we suspect that one type has such a cost function and can estimate the smallest marginal cost for this type over all waiting times t, then our result can provide a simple sufficiency condition for the optimality of prioritizing

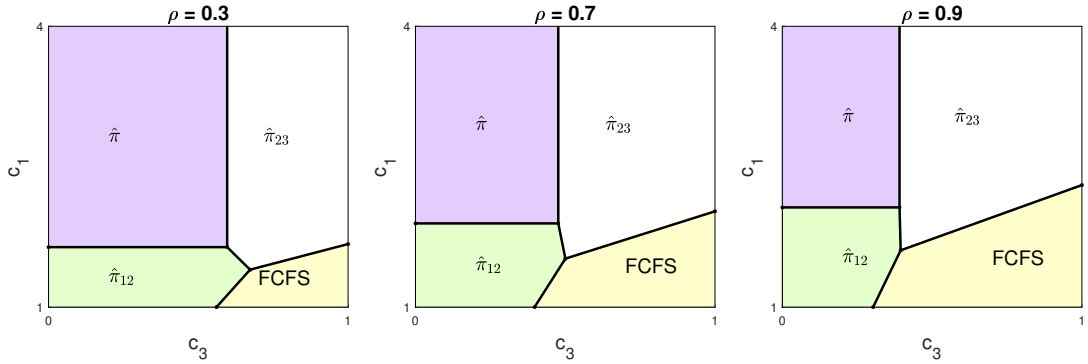

Figure 3 **Optimal policy in Π^F with $K=3$, $p_i=1/3$, $C_i(t)=c_it^2$ for $i\in\{1,2,3\}$, $c_2=1$, $\rho\in\{0.3,0.7,0.9\}$, and exponentially distributed service times with mean one.**

this type over all other types with linear costs. As an example, suppose that $C_1(t)=h_3t^3-h_2t^2+h_1t$ for $t\geq 0$, where h_1,h_2,h_3 are positive constants such that $C_1'(t)>0$, i.e., $h_2^2<3h_1h_3$. (It is easy to see that this cubic cost function is S-shaped.) Then, our result implies that type 1 should receive the highest priority if $\min_{t\geq 0}C_1'(t)=h_1-h_2^2/(3h_3)>\max_{j=2,\ldots,K}\{c_j\mu_j\}/\mu_1$.

7. Numerical study

To obtain a better understanding of priority assignment in systems with multiple types of customers, we conducted a numerical study on a system with three types as presented in Section 7.1. We also conducted an exploratory numerical analysis into when it would be desirable to consider more complex policies that take into account the waiting times of customers; see Section 7.2.

7.1. Fixed priority policies for systems with three types of customers

Consider a system with $K=3$, where service times are i.i.d. exponentially distributed with rate $\mu=1$ for all customers types, and the cost function for type i customers is in the form of $C_i(t)=c_it^2$ for $i=1,2,3$. (For notational convenience, we drop the superscript of $c_i^{(2)}$ throughout Section 7.) We fix c_2 to be one and without loss of generality, we set the cost parameters such that $c_1>c_2>c_3$.

Recall that there are thirteen policies in Π^F when $K=3$ as discussed at the beginning of Section 4. Instead of computing and comparing the long-run average costs for all thirteen policies to identify the best policy π^* in Π^F, we first eliminate some of these policies using Corollary 1 and the assumption that $c_1>c_2>c_3$. In particular, we can eliminate policies that give priority to types with smaller values of c_i; see the discussion in the paragraph following Corollary 1. Hence, out of the six policies in Π^F that assign a different priority order to each type, we only need to consider the one that assigns priority order i to type i customers for $i=1,2,3$. Furthermore, for any policy in Π^F that pools types k and ℓ for $1\leq k<\ell\leq 3$, the waiting cost function for the pooled group of customers will be $C_{k\ell}(t)=c_{k\ell}t^2$, $t\geq 0$, where $c_{k\ell}=(p_kc_k+p_\ell c_\ell)/(p_k+p_\ell)$. Then, since $c_{12}>c_3$ and $c_1>c_{23}$, we can eliminate the policy that prioritizes type 3 over the pooled group of types 1 and 2 as well as the policy that prioritizes the pooled group of types 2 and 3 over type 1.

After eliminating seven policies from Π^F, we numerically compute and compare the long-run average costs for the remaining six policies to find π^* for different values of $c_1>1>c_3$ when the proportion of each type p_i is $1/3$ and the traffic intensity is $\rho\in\{0.3,0.7,0.9\}$. For these parameters, our numerical results show that π^* can only be one of the following four policies: (i) policy $\hat{\pi}$ that assigns priority order i to type i customers for $i=1,2,3$; (ii) policy $\hat{\pi}_{12}$ that prioritizes the pooled group of types 1 and 2 customers over type 3; (iii) policy $\hat{\pi}_{23}$ that prioritizes type 1 customers over the pooled group of types 2 and 3; and (iv) FCFS. The two policies that pool types 1 and 3 and prioritizes either the pooled group or type 2 were never the best under all tested parameters. (The comparisons for distinct values of p_i's are similar and are provided in Appendix A.4.)

Corollary 1 provides the following partial comparison of the remaining four policies:

(i) For $k \in \{1,2\}$ and $\ell = k+1$, $C_{\hat{\pi}} \le C_{\hat{\pi}_{k\ell}}$ if and only if $\dfrac{c_k}{c_{k+1}} \ge \Gamma_k \equiv 1 + \dfrac{\bar{\rho}_{k-1}\left(\bar{\rho}_{k+1}^{-1} + \bar{\rho}_k^{-1} - 1\right)}{\bar{\rho}_k\left(\bar{\rho}_{k+1}^{-1} + \bar{\rho}_k^{-1} + \bar{\rho}_{k-1}^{-1} - 1\right)}$;

(ii) $C_{\mathrm{FCFS}} \le C_{\hat{\pi}_{12}}$ if and only if $\dfrac{c_{12}}{c_3} \le \Gamma_{12} \equiv \dfrac{2-\rho_1-\rho_2+(1-\rho)(1-\rho_1-\rho_2)^{-1}}{2-\rho_1-\rho_2-\rho}$; and

(iii) $C_{\mathrm{FCFS}} \le C_{\hat{\pi}_{23}}$ if and only if $\dfrac{c_1}{c_{23}} \le \Gamma_{23} = \dfrac{2-\rho_1+(1-\rho)(1-\rho_1)^{-1}}{2-\rho_1-\rho}$.

Figure 3 provides plots of optimality regions determined partially by bounds $\Gamma_1, \Gamma_2, \Gamma_{12}$, and Γ_{23} provided above. More specifically, the upper left-most corner is where $\hat{\pi}$ is better than $\hat{\pi}_{12}$ and $\hat{\pi}_{23}$ (determined by lines $c_1 = \Gamma_1$ and $c_3 = \Gamma_2^{-1}$, respectively) and lower right-most corner is where FCFS is better than $\hat{\pi}_{12}$ and $\hat{\pi}_{23}$ (determined by lines $c_1 = \Gamma_{12}(1 + p_2/p_1)c_3 - p_2/p_1$ and $c_1 = \Gamma_{23}(p_2 + p_3 c_3)/(p_2 + p_3)$, respectively). By using numerical comparisons, we also find that $\hat{\pi}$ and FCFS are indeed the best in Π_F in these respective regions. The optimality regions for $\hat{\pi}_{12}$ and $\hat{\pi}_{23}$ are found numerically in Figure 3. (Our analytical results do not provide comparisons between FCFS and $\hat{\pi}$ or between $\hat{\pi}_{12}$ and $\hat{\pi}_{23}$.) We make the following observations from Figure 3, and Figures A.1 and A.2 in Appendix A.4, which extend most of our analytical observations for the case with two types of customers in Section 5.1 to the case with three types:

(1) Assigning different priority orders to individual types (policy $\hat{\pi}$) is the best if the cost coefficients for all types are significantly different, i.e., $c_1/c_2 \ge \Gamma_1 > 1 > 1/\Gamma_2 \ge c_3/c_2$.
(2) FCFS is the best when the cost coefficients of all three types are close to each other, i.e, c_1/c_2 and c_3/c_2 are both close to 1.
(3) As λ (and hence ρ) increases, the region where assigning individual priority is the best shrinks, and the region where FCFS is the best enlarges.
(4) Pooling types k and $k+1$ is better than assigning individual priority order or FCFS if the cost coefficients of the pooled types are close but are significantly different from that of the remaining type. For example, if $c_1/c_2 \le \Gamma_1$, and c_3 is sufficiently small, then pooling type 1 and type 2 together and prioritizing the group over type 3 is the best (the lower-left corner).
(5) The optimality region for policies that prioritize a particular type (or group) over the others shrinks if the proportion of that type (or group) increases. For example, if p_1 increases, then the regions where type 1 should be prioritized over the remaining customers (i.e., the optimality regions for policies $\hat{\pi}$ and $\hat{\pi}_{23}$) shrink; and if p_2 increases (for fixed p_1), then the optimality region of policy $\hat{\pi}$ becomes smaller.
(6) A proportionally dominant type (or group) should be prioritized under heavy traffic only if its cost coefficient is much larger than that of the remaining type(s).

7.2. Comparison of fixed priority policies with a dynamic priority policy (G-$c\mu$ rule)

In this section, we numerically compare the performance of the best policy within Π^F with the performance of a well-known dynamic policy that takes into account the waiting times of customers for prioritization. In particular, we compare the best policy π^* in Π^F with the generalized $c\mu$ (G-$c\mu$) rule under a wide range of parameter settings. Our goal is to identify conditions under which it would be worthwhile to use the G-$c\mu$ rule as opposed to π^* and also conditions under which the additional complexity of the G-$c\mu$ rule does not bring much benefits. (Recall that G-$c\mu$ rule gives priority to the customer with the largest $C_i'(t)\mu_i$ value.) The comparison was made with the G-$c\mu$ rule, a heuristic, because we were not able to identify the optimal dynamic policy within the set of all policies that take into account waiting times due to a large state space. Note that the G-$c\mu$ rule is optimal for convex cost functions under heavy traffic (Van Mieghem 1995).

Here, we present our results for systems with $K = 2$, but note that our study on systems with three types yields similar conclusions (see Appendix A.4). Assume that service times for type $i \in \{1,2\}$ customers are exponentially distributed with rate μ_i, where μ_2 is fixed at one per unit time, and the cost functions take the form $C_1(t) = c_1 t^2$ and $C_2(t) = t^2$, $t \ge 0$. We consider 81 different scenarios corresponding to all combinations of $\rho \in \{0.3, 0.7, 0.9\}$, $p_1 \in \{0.1, 0.5, 0.9\}$, $\mu_1 \in \{0.2, 1, 5\}$,

and $c_1 \in \{0.1, 0.9, 5\}$. We can identify the best fixed priority policy in Π^F by computing the values of R^{π_1} and R^{π_2} in Corollary 1 (reported in Table A.4 in the Appendix): π_2 has the smallest cost if $c_1\mu_1 < R^{\pi_2}$; π_1 has the smallest cost if $c_1\mu_1 > R^{\pi_1}$; and FCFS has the smallest cost if $R^{\pi_2} \leq c_1\mu_1 \leq R^{\pi_1}$. To obtain the long-run average cost under the G-$c\mu$ rule (denoted by C_G), we built a simulation model (using Simio 11 simulation software), where we computed a priority index for each customer in the queue and assigned non-preemptive priority to the one with the largest index. Under the specific cost structure and experimental setting of this section, the priority index for a type i customer who waited for $t \geq 0$ time units, and determined by the G-$c\mu$ rule, is given by $2c_i t\mu_i$. We ran 100 independent replications of length 60,000 minutes for each scenario and truncated the first 6,000 minutes based on a warm-up period analysis. We report the mean relative cost difference by using the G-$c\mu$ rule over the best fixed priority policy, i.e., $(C_G - C_{\pi^*}) \times 100/C_{\pi^*}$ (in percentage) and a 95% confidence interval (C.I.) on this relative cost difference from the simulation runs. If the C.I. does not contain zero, then we conclude that there is statistical evidence that π^* and the G-$c\mu$ rule are different, where the comparison is in favor of π^* for a positive C.I. and the G-$c\mu$ rule for a negative one. Figure 4 presents these results.

From these simulation results, we find that the cost difference between the best fixed priority policy and the G-$c\mu$ rule is insignificant in most scenarios, especially when the traffic intensity is light or moderate, and the G-$c\mu$ rule may have smaller costs in scenarios under heavy traffic. In several scenarios, e.g., $c_1 = 5, \mu_1 = 5, p_1 = 0.9$ with traffic intensity 0.7 or 0.9 (where $\pi^* = \pi_1$), and $c_1 = 0.1, \mu_1 = 0.2, p_1 = 0.1$ with traffic intensity 0.7 (where $\pi^* = \pi_2$), the best fixed priority policy performs better than the G-$c\mu$ rule. In these scenarios, we find that the prioritized type has significantly higher proportion, cost coefficient, and service rate. Furthermore, we notice that when FCFS is the best fixed priority policy, either its performance is similar to that of the G-$c\mu$ rule or the G-$c\mu$ rule outperforms it. We also observe that for heavy-traffic scenarios, where the parameters fall close to the thresholds that characterize the optimal fixed priority policy reported in Table A.4 (possibly suggesting that none of the fixed priority policies stands out), the G-$c\mu$ rule performs better than the best fixed priority policy. Hence, it would be worthwhile to consider the more complex G-$c\mu$ rule over a fixed priority policy when the traffic is heavy and there is not a clearly more "important" type. One could assess whether there is clearly a more important type or not by considering how far the system parameters land from the thresholds of the best fixed priority policy. If they are closer to a threshold, such as in scenarios $c_1 = 5, \mu_1 = 1, p_1 = 0.9, \rho = 0.9$ or $c_1 = 0.9, \mu_1 = 0.2, p_1 = 0.1, \rho = 0.9$, then this could be taken as an indicator that there is not a clearly more important type and hence G-$c\mu$ rule should be considered. On the other hand, when the traffic is light or the system parameters fall farther away from the thresholds, e.g., when one type has a substantially larger cost, service rate, and proportion, then it is not necessary to use the G-$c\mu$ rule and in fact it could be better to use the best fixed priority policy, which does not require knowing the cost function precisely and is much simpler to implement.

8. Conclusions

In order to answer some fundamental questions surrounding prioritization of certain customer groups in a service system, we studied a single-server queueing model with stationary Poisson arrivals of multiple types of customers with possibly distinct service time distributions and nonlinear waiting cost functions. When waiting costs are nonlinear functions of time, it is known that in general, the priority policy that minimizes the long-run average waiting costs is dynamic, i.e., dependent on the durations of time customers in the queue have already spent waiting, in addition to their types. However, in practice, the most commonly employed policies are still first-come-first-served (FCFS) and strict fixed priority policies that give exclusive priority to one of the types of customers independently of the system state. In this paper, we compared these fixed priority policies (including FCFS) in terms of their long-run average performance and derived several managerial insights by focusing mostly on the case with quadratic waiting costs.

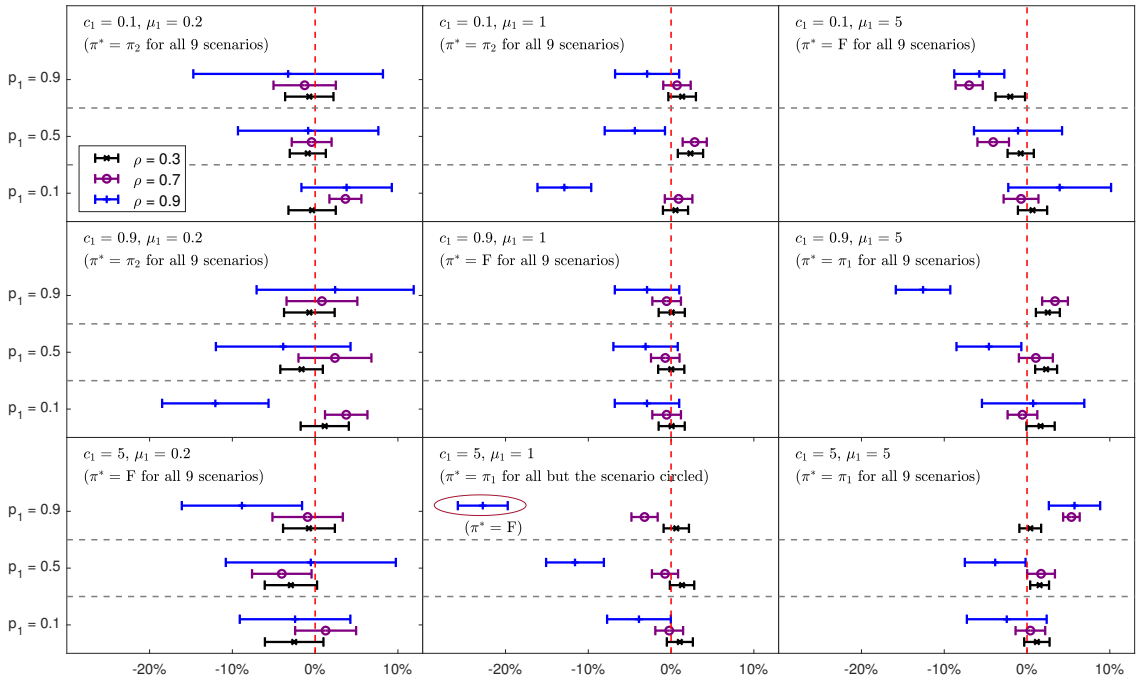

Figure 4 **95% C.I. of the relative cost difference between G-$c\mu$ rule and π^* for the case with $K = 2$, where negative values indicate that G-$c\mu$ rule has a smaller cost than π^*.**

It is well known that if all customers have linear waiting costs, then only the product of the rates of service and waiting cost will affect the characterization of optimal policies, and the higher the number of priority classes the better it is – FCFS being the worst. However, this is no longer the case when cost functions are nonlinear. More specifically, for quadratic cost functions, we concluded that splitting the customer population into as many priority classes as possible may actually increase the long-run average waiting costs if one were to apply only fixed priority policies. Fixed priority policies can perform better than FCFS when there is sufficient *heterogeneity* in the population, and hence, the benefit gained by prioritizing one group over the other compensates for the damage caused by lowering the priority of the rest. For quadratic costs, we found that the heterogeneity of the population is determined by the population mix (i.e., proportions of each customer type in the population) and first three moments of service times as well as cost parameters. Furthermore, we showed that the arrival rate has a direct effect on the decision to prioritize or not to prioritize. Specifically, we observed that the parameter region where FCFS is best enlarges as the arrival rate increases. Hence, haphazardly replacing FCFS discipline with a strict fixed priority policy without considering system parameters such as traffic intensity and service-time variability may lead to inferior system performance when there is any concern that the waiting cost functions might not be linear. We found that one should be especially cautious with prioritizing types that constitute a large proportion of the population mix. One setting where it would be safe to replace FCFS discipline with a strict fixed priority policy is when one type has a quadratic cost function and the derivative of the cost function of all other customers is bounded from above (as in a linear cost function). In such a case, a policy that prioritizes the type with quadratic cost is better under heavy traffic regardless of the service time distributions.

As a byproduct of our study on quadratic costs, we also obtained some useful results on the problem of minimizing the variance of steady-state waiting times, which is widely accepted to be equivalent to maximizing fairness in queueing systems. Earlier work showed that FCFS is the fairest policy when the population is homogenous in terms of service time variability. For a population with heterogenous service time variability, we showed that FCFS is still the fairest policy if the traffic intensity is sufficiently large and no type is significantly dominant in numbers. In particular,

one of our results imply that FCFS is better than any policy that groups customer types into two priority classes if the traffic intensity is larger than $1/(1+\sqrt{\underline{p}})$, where $\underline{p}$ is the smallest proportion of any type in the population. However, if the traffic intensity is not heavy, then we showed that prioritizing the type with smaller service-time variance could actually be fairer than FCFS.

Since the focus of this work was on the use (or misuse) of simple but popular fixed priority policies in systems with non-linear waiting costs, we mostly excluded more complex priority policies such as those that use waiting time information of customers while giving priority decisions. Perhaps the most well-known policy in this set is the G-$c\mu$ rule, which is shown to be optimal under heavy traffic and convex waiting costs. An important future research direction would be to study conditions under which it would be better to use these more complex policies. In this paper, we provided an exploratory analysis to facilitate interest on this research question by conducting a simulation study that compares the best fixed priority policy with the G-$c\mu$ rule under quadratic waiting costs. We found that the G-$c\mu$ rule performs better than the best fixed priority policy for a heavily loaded system when the customer population is not sufficiently heterogeneous. On the other hand, when the traffic is not heavy, or one type has substantially larger cost of waiting, service rate, and proportion of the demand, then the best fixed priority policy, which is much easier to implement and does not require precise knowledge on the waiting cost function, performs similarly or even slightly better than the G-$c\mu$ rule. We believe that more research is needed to support these claims especially given that the G-$c\mu$ rule is not always an optimal dynamic policy. Studying randomized priority policies, where the priority is not fixed but randomly assigned to different types, would be an interesting and useful future research direction as well because these policies would be a good compromise between simple fixed priority policies and more complex state-dependent ones.

Another interesting future research direction would be to study the same problem under other specific waiting-cost structures in more detail such as cost functions with exponential or concave growth. For some preliminary results in this direction, we refer interested readers to Section 2.7 in Ouyang (2016).

References

Ansell P, Glazebrook KD, Niño-Mora J, O'Keeffe M (2003) Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research* 57(1):21–39.

Argon NT, Ding L, Glazebrook KD, Ziya S (2009) Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences* 23(02):175–203.

Argon NT, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* 11(4):674–693.

Ata B, Peng X (2018) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research* 66(1):163–183.

Ata B, Tongarlak MH (2013) On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems* 74(1):65–104.

Avi-Itzhak B, Levy H (2004) On measuring fairness in queues. *Advances in Applied Probability* 36(03):919–936.

Bertsimas D (1995) The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Systems* 21:337–389.

Bispo CF (2013) The single-server scheduling problem with convex costs. *Queueing Systems* 73(3):261–294.

Budhiraja A, Ghosh A, Liu X (2014) Scheduling control for markov-modulated single-server multiclass queueing systems in heavy traffic. *Queueing Systems* 78(1):57–97.

Cobham A (1954) Priority assignment in waiting line problems. *Journal of the Operations Research Society of America* 2(1):70–76.

Cox DR, Smith WL (1961) *Queues* (Methuen).

Di Crescenzo A (1999) A probabilistic analogue of the mean value theorem and its applications to reliability theory. *Journal of Applied Probability* 36(03):706–719.

Ding Y, Park E, Nagarajan M, Grafstein E (2019) Patient prioritization in emergency department triage systems: An empirical study of the canadian triage and acuity scale (ctas). *Manufacturing & Service Operations Management* 21(4):723–741.

El-Taha M, Stidham Jr S (1999) *Sample-Path Analysis of Queueing Systems* (Springer Science & Business Media).

Ghahramani S, Wolff RW (1989) A new proof of finite moment conditions for GI/G/1 busy periods. *Queueing Systems* 4(2):171–178.

Glazebrook KD, Lumley R, Ansell P (2003) Index heuristics for multiclass M/G/1 systems with nonpreemptive service and convex holding costs. *Queueing Systems* 45(2):81–111.

Gross D, Shortle JF, Thompson JM, Harris C (2008) *Fundamentals of Queueing Theory* (John Wiley & Sons), Fourth edition.

Gross OA (1962) Preferential arrangements. *The American Mathematical Monthly* 69(1):4–8.

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

Haji R, Newell GF (1971) Optimal strategies for priority queues with nonlinear costs of delay. *SIAM Journal on Applied Mathematics* 20(2):224–240.

Harrison JM (1975) Dynamic scheduling of a multiclass queue: Discount optimality. *Operations Research* 23(2):270–282.

Jaiswal NK (1968) *Priority Queues* (Academic Press).

Jenkins J, McCarthy LM, Sauer LM, Green SB, Stuart S, Thomas T, Hsu E (2008) Mass-casualty triage: time for an evidence-based approach. *Prehospital and Disaster Medicine* 23(1):3–8.

Jiang Y, Tham CK, Ko CC (2002) Delay analysis of a probabilistic priority discipline. *European transactions on telecommunications* 13(6):563–577.

Kakalik J, Little J (1971) Optimal service policy for the M/G/1 queue with multiple classes of arrivals. Technical report, Rand Corporation Report.

Katayama T, Takahashi Y (1992) Analysis of a two-class priority queue with bernoulli schedules. *Journal of the Operations Research Society of Japan* 35(3):236–249.

Kingman J (1962) The effect of queue discipline on waiting time variance. *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 58, 163–164 (Cambridge University Press).

Kleinrock L (1965) A conservation law for a wide class of queueing disciplines. *Naval Research Logistics Quarterly* 12(2):181–192.

Klimov G (1974) Time-sharing service systems I. *Theory of Probability & Its Applications* 19(3):532–551.

Klimov G (1979) Time-sharing service systems. II. *Theory of Probability & Its Applications* 23(2):314–321.

Kulkarni V (2009) *Modeling and Analysis of Stochastic Systems* (CRC Press), Second edition.

Larranaga M, Ayesta U, Verloop IM (2015) Asymptotically optimal index policies for an abandonment queue with convex holding cost. *Queueing Systems* 81(2-3):99–169.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research* 52(6):836–855.

Miller DR (1960) Priority queues. *The Annals of Mathematical Statistics* 31(1):86–103.

Nain P (1989) Interchange arguments for classical scheduling problems in queues. *Systems & Control Letters* 12(2):177–184.

Ouyang H (2016) *Prioritization In Service Systems With Nonlinear Delay Costs*. Ph.D. thesis, University of North Carolina at Chapel Hill.

Parlar M, Sharafali M (2014) Optimal design of multi-server markovian queues with polynomial waiting and service costs. *Applied Stochastic Models in Business and Industry* 30(4):429–443.

Pinedo M (1983) Stochastic scheduling with release dates and due dates. *Operations Research* 31(3):559–572.

Roman S (1980) The formula of Faa di Bruno. *American Mathematical Monthly* 87(10):805–809.

Sacco WJ, Navin DM, Fiedler KE, Waddell I, Robert K, Long WB, Buckman RF (2005) Precise formulation and evidence-based application of resource-constrained triage. *Academic Emergency Medicine* 12(8):759–770.

Shaked M, Shanthikumar J (2007) *Stochastic Orders* (Springer Science & Business Media).

Shanthikumar JG, Yao DD (1992) Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research* 40(3-supplement-2):S293–S299.

Sun Z, Argon NT, Ziya S (2017) Patient triage and prioritization under austere conditions. *Management Science* 64(10):4471–4489.

Takács L (1964) Priority queues. *Operations Research* 12(1):63–74.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized cμ rule. *The Annals of Applied Probability* 5(3):809–833.

Vasicek OA (1977) An inequality for the variance of waiting time under a general queuing discipline. *Operations Research* 25(5):879–884.

Wolff RW (1989) *Stochastic Modeling and the Theory of Queues* (Pearson College Division).

Appendix

In this Appendix, we provide proofs of theoretical results and other supplemental material.

A.1. Notation, definitions, and lemmas

We first provide three tables of notation used in the main paper and the Appendix.

Table A.1 Notation for System Parameters

λ	arrival rate of all customers
p_i	probability that an arriving customer belongs to type i
μ_i	service rate of type i customers
$\tau_i^{(n)}$	nth moment of service time of type i customers for $n \geq 2$
ρ_i	traffic intensity for type i customers, i.e., $\lambda p_i/\mu_i$
$\bar{\rho}_k$	$1-\sum_{j=1}^{k}\rho_j$ for $k=1,\ldots,K$ and $\bar{\rho}_0=1$
ρ	traffic intensity for all customers, i.e., $\sum_{j=1}^{K}\rho_j$
μ	service rate for a random customer, i.e., $(\sum_{j=1}^{K} p_j/\mu_j)^{-1}$
$\tau^{(n)}$	nth moment of service time for a random customer, i.e., $\sum_{j=1}^{K} p_j\tau_j^{(n)}$, for $n \geq 2$
$C_i(\cdot)$	waiting cost function for type i customers
C_γ	long-run average waiting cost under policy γ
K^γ	number of priority orders under policy γ
$\gamma(i)$	priority order of type i customers under policy γ, where $\gamma(i) \in \{1,2,\ldots,K^\gamma\}$
$p_{[j]}^\gamma$	probability that an arriving customer has priority order j under policy γ, i.e., $\sum_{\{i:\gamma(i)=j\}} p_i$
$\mu_{[j]}^\gamma$	service rate of customers with priority order j under policy γ
$\tau_{[j]}^{(n),\gamma}$	nth moment of service time of customers with priority order j under policy γ for $n \geq 2$
$\rho_{[j]}^\gamma$	traffic intensity for customers with priority order j under policy γ
$\bar{\rho}_{[k]}^\gamma$	$1-\sum_{j=1}^{k}\rho_{[j]}^\gamma$ for $k=1,\ldots,K^\gamma$ and $\bar{\rho}_0^\gamma=1$
$\Lambda_{\leq k}^\gamma$	arrival rate of customers with priority order k or smaller under policy γ, i.e., $\lambda\sum_{j=1}^{k} p_{[j]}^\gamma$

Table A.2 Notation for random variables (RV), their cumulative density functions (CDF), and Laplace-Stieltjes transforms (LST)

RV	CDF	LST	Definition
$V_{i,k}^{\gamma,x_0}$	-	-	waiting time of the kth arriving type i customer under policy γ and initial state x_0
W_i^γ	-	$\widetilde{W}_i^\gamma(\cdot)$	steady-state waiting time of a type i customer under policy γ
$W^{\bar{\pi}_k}$	-	-	steady-state waiting time of types k and $k+1$ customers under policy $\bar{\pi}_k$
$W_{[j]}^\gamma$	-	$\widetilde{W}_{[j]}^\gamma(\cdot)$	steady-state waiting time for a customer with priority order j under policy γ
-	$S_i(\cdot)$	$\tilde{S}_i(\cdot)$	service time for type i customers
-	$S_{[j]}^\gamma(\cdot)$	$\tilde{S}_{[j]}^\gamma(\cdot)$	service time for customers with priority order j under policy γ, i.e., $S_{[j]}^\gamma(x) \equiv \sum_{\{i:\gamma(i)=j\}} p_i S_i(x)/p_{[j]}^\gamma$
-	$S_{\leq k}^\gamma(\cdot)$	$\tilde{S}_{\leq k}^\gamma(\cdot)$	service time for customers with priority order k or smaller under policy γ, i.e., $S_{\leq k}^\gamma(x) \equiv \sum_{j=1}^k p_{[j]}^\gamma S_{[j]}^\gamma(x) / \sum_{j=1}^k p_{[j]}^\gamma$
-	$S_{>k}^\gamma(\cdot)$	$\tilde{S}_{>k}^\gamma(\cdot)$	service time for customers with priority order greater than k under policy γ, i.e., $S_{>k}^\gamma(x) \equiv \sum_{j=k+1}^{K^\gamma} p_{[j]}^\gamma S_{[j]}^\gamma(x) / \sum_{j=k+1}^{K^\gamma} p_{[j]}^\gamma$
S	$S(\cdot)$	$\tilde{S}(\cdot)$	service time for a random customer, i.e., $S(x) \equiv \sum_{j=1}^K p_j S_j(x)$
-	$B_{p,j}^\gamma(\cdot)$	-	length of a busy period in an M/G/1 queue with arrival rate $\Lambda_{\leq p-1}^\gamma$ and service time distribution $S_{\leq p-1}^\gamma(x)$ for $p = 2, 3, \ldots, K^\gamma$ under policy γ in which there are $j \geq 1$ customers initially in the system
$U_i^{\gamma_1,\gamma_2}$	-	-	$\Psi(W_i^{\gamma_1}, W_i^{\gamma_2})$ for $i \in \{k, k+1\}$, $\gamma_1, \gamma_2 \in \{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, and $\gamma_1 \neq \gamma_2$ where $\Psi(\cdot)$ is defined in Definition A.1
U_i^γ	-	-	$U_i^{\gamma,\bar{\pi}_k}$ for $\gamma \in \{\pi_k, \pi_{k+1}\}$ and $i \in \{k, k+1\}$
$T_{B_{\neq i}}$	-	$\tilde{B}_{\neq i}(\cdot)$	length of a busy period where only type $j \in \{1, 2, \ldots, K\} \setminus \{i\}$ customers arrive
-	-	$\tilde{B}(\cdot)$	length of a busy period starting from an empty and idle system for an M/G/1 queue under any policy in Π

Table A.3 Notation for Policy Parameters

Proposition 2	$\delta_i^{\gamma_1,\gamma_2} \equiv E\left[C_i'\left(U_i^{\gamma_1,\gamma_2}\right)\right]$, for $i \in \{k, k+1\}$, $\gamma_1, \gamma_2 \in \{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, and $\gamma_1 \neq \gamma_2$
	$\delta_i^\gamma \equiv \delta_i^{\gamma,\bar{\pi}_k}$, for $i \in \{k, k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$
Proposition 3	$M_i^\gamma \equiv \dfrac{E\left[\left(W^{\bar{\pi}_k}\right)^2\right] - E\left[\left(W_i^\gamma\right)^2\right]}{E\left[W^{\bar{\pi}_k}\right] - E\left[W_i^\gamma\right]}$, for $i \in \{k, k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$
Corollary 1	$R^\gamma \equiv M_{k+1}^\gamma / M_k^\gamma$ for $\gamma \in \{\pi_k, \pi_{k+1}\}$

Next, we introduce several definitions and lemmas that will be used in the proofs of our results.

LEMMA A.1. *For policies π_k, π_{k+1}, and $\bar{\pi}_k$ that are defined in Section 4, we have:*
(a) $W_j^{\pi_k} =_{st} W_j^{\pi_{k+1}} =_{st} W_j^{\bar{\pi}_k}$ *for* $j \in \{1,2,\ldots,K\} \setminus \{k,k+1\}$, *where* $=_{st}$ *means equivalence in distribution;*
(b) $W_i^{\pi_i} \leq_{st} W^{\bar{\pi}_k} \leq_{st} W_{2k+1-i}^{\pi_i}$ *for* $i \in \{k,k+1\}$, *where* $\leq_{st}$ *denotes usual stochastic ordering, see e.g., Section 1.A.1 of Shaked and Shanthikumar (2007); and*
(c) $E\left[W^{\bar{\pi}_k}\right] = \dfrac{\lambda\tau^{(2)}}{2\bar{\rho}_{k+1}\bar{\rho}_{k-1}}, E\left[W_i^{\pi_i}\right] = \dfrac{\lambda\tau^{(2)}}{2(\bar{\rho}_{k-1}-\rho_i)\bar{\rho}_{k-1}}, E\left[W_{2k+1-i}^{\pi_i}\right] = \dfrac{\lambda\tau^{(2)}}{2\bar{\rho}_{k+1}(\bar{\rho}_{k-1}-\rho_i)}$, *for* $i \in \{k,k+1\}$.

Proof of Lemma A.1: (*a*) The distribution function for the steady-state waiting times of customers with priority order p under policy γ is given by Equation (17) in Takács (1964) as follows:

$$P\{W_{[p]}^{\gamma} \leq x\} = \int_0^x \left[\sum_{j=0}^{\infty} e^{-\Lambda_{\leq p-1}^{\gamma} y} \frac{(\Lambda_{\leq p-1}^{\gamma} y)^j}{j!} B_{p,j}^{\gamma}(x-y)\right] dW_{\leq p}^{*\gamma}(y), \tag{EC.1}$$

where $B_{p,j}^{\gamma}(x)$ for $p = 2,3,\ldots,K^{\gamma}$ is the distribution function of the length of a busy period in an M/G/1 queue with arrival rate $\Lambda_{\leq p-1}^{\gamma}$ and service time distribution $S_{\leq p-1}^{\gamma}(x)$ in which there are $j \geq 0$ customers initially in the system, $B_{1,j}^{\gamma}(x) = 1$ for all $x \geq 0$ and $j \geq 0$, and $W_{\leq p}^{*\gamma}(y)$ for $p = 1,2,\ldots,K^{\gamma}$ is the distribution function of the steady-state waiting time for a customer with priority order $\leq p$ under a modified policy, defined as follows: customers of priority order $\leq p$ under γ are pooled and served according to FCFS regardless of their actual priority order under policy γ and their service times are i.i.d. with distribution $S_{\leq p}^{\gamma}(x)$, while all other customers are served according to γ.

For $p = 1,2,\ldots,k-1$, it is easy to see that $\Lambda_{\leq p-1}^{\pi_k} = \Lambda_{\leq p-1}^{\pi_{k+1}} = \Lambda_{\leq p-1}^{\bar{\pi}_k}$, $S_{\leq p-1}^{\pi_k}(x) = S_{\leq p-1}^{\pi_{k+1}}(x) = S_{\leq p-1}^{\bar{\pi}_k}(x)$, and hence $B_{p,j}^{\pi_k}(x) = B_{p,j}^{\pi_{k+1}}(x) = B_{p,j}^{\bar{\pi}_k}(x)$ for all $x \geq 0$ and $j \geq 0$, and $W_{\leq p}^{*\pi_k}(y) = W_{\leq p}^{*\pi_{k+1}}(y) = W_{\leq p}^{*\bar{\pi}_k}(y)$ for all $y \geq 0$. Similarly, for $p = k+2,\ldots,K$, we have $\Lambda_{\leq p-1}^{\pi_k} = \Lambda_{\leq p-1}^{\pi_{k+1}} = \Lambda_{\leq p-2}^{\bar{\pi}_k}$, $S_{\leq p-1}^{\pi_k}(x) = S_{\leq p-1}^{\pi_{k+1}}(x) = S_{\leq p-2}^{\bar{\pi}_k}(x)$, and $B_{p,j}^{\pi_k}(x) = B_{p,j}^{\pi_{k+1}}(x) = B_{p-1,j}^{\bar{\pi}_k}(x)$ for all $x \geq 0$ and $j \geq 0$, and $W_{\leq p}^{*\pi_k}(y) = W_{\leq p}^{*\pi_{k+1}}(y) = W_{\leq p-1}^{*\bar{\pi}_k}(y)$ for all $y \geq 0$. Thus, by (EC.1), $W_p^{\pi_k} =_{st} W_p^{\bar{\pi}_k} =_{st} W_p^{\pi_{k+1}}$ for $p \in \{1,2,\ldots,K\} \setminus \{k,k+1\}$.

(*b*) We use sample path arguments to prove the stochastic inequalities. First, we fix $i \in \{k,k+1\}$. In this proof, type k and type $k+1$ customers (who have priority order k and $k+1$ under policy $\pi = \pi_k$) will be referred to as relevant customers, type i customers will be called higher priority customers, and type i' customers will be called lower priority customers under policy π_i, where $i' = 2k+1-i$.

We index the relevant customers by their arrival order to the system, and let s_j be the arriving time of the jth relevant customer. Then, for the ℓth and jth relevant customers, where $j > \ell \geq 1$, we have $s_j > s_l$. Let t_j^{γ} be the service starting time of the jth relevant customer under policy $\gamma \in \{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, then $t_j^{\gamma} \geq s_j$. Let also V_j^{γ} denote the waiting time of the jth relevant customer under policy γ, then $V_j^{\gamma} = t_j^{\gamma} - s_j$ for $j = 1,2,\ldots$.

Under $\bar{\pi}_k$, we have $t_1^{\bar{\pi}_k} < t_2^{\bar{\pi}_k} < \cdots$ with probability one. Let j_1 be the index of the first lower priority customer whose service starts when there are higher priority customers waiting, and j_2 be the index of the first higher priority customer in the queue when j_1 starts service under $\bar{\pi}_k$. Then, the customers indexed from j_1 to $j_2 - 1$ are all lower priority customers. Note that $s_{j_1} < \cdots < s_{j_2-1} < s_{j_2} < t_{j_1}^{\bar{\pi}_k} < \cdots < t_{j_2-1}^{\bar{\pi}_k} < t_{j_2}^{\bar{\pi}_k}$.

Consider a policy π' that follows $\bar{\pi}_k$ except that it serves customer j_2 before it serves lower priority customers $j_1,\ldots,j_2-1$. For customer j_2, who is a higher priority customer, $t_{j_2}^{\pi'} = t_{j_1}^{\bar{\pi}_k} < t_{j_2}^{\bar{\pi}_k}$ and $V_{j_2}^{\pi'} = t_{j_2}^{\pi'} - s_{j_2} < t_{j_2}^{\bar{\pi}_k} - s_{j_2} = V_{j_2}^{\bar{\pi}_k}$. For $\ell = j_1,\ldots,j_2-1$, who are all lower priority customers, $t_l^{\pi'} > t_l^{\bar{\pi}_k}$ and $V_l^{\pi'} = t_l^{\pi'} - s_l > t_l^{\bar{\pi}_k} - s_l = V_l^{\bar{\pi}_k}$. For any $\ell \notin \{j_1,\ldots,j_2\}$, we have $V_l^{\pi'} = V_l^{\bar{\pi}_k}$.

If we keep changing the service order like this when there are lower priority customers starting service while higher priority customers are waiting in the queue, then we will eventually reach policy π_i. This coupling argument then will yield $V_{i,n}^{\pi_i,x_0} \leq_{st} V_{i,n}^{\bar{\pi}_k,x_0}$ and $V_{i,n}^{\pi_i,x_0} \geq_{st} V_{i',n}^{\bar{\pi}_k,x_0}$ for $n \geq 1$. Since W_i^{γ} is the steady-state waiting time for type i customers under policy γ, then, as $n \to \infty$, $V_{i,n}^{\gamma,x_0} \xrightarrow{d} W_i^{\gamma}$ and $V_{i',n}^{\gamma,x_0} \xrightarrow{d} W_{i'}^{\gamma}$ for $\gamma \in \{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, where $\xrightarrow{d}$ denotes convergence in distribution, and hence, according to Theorem 1.A.3(c) in Shaked and Shanthikumar (2007), we have $W_i^{\pi_i} \leq_{st} W^{\bar{\pi}_k} \leq_{st} W_{i'}^{\pi_i}$.

(*c*) The result follows directly from Equation (68) of Takács (1964). □

DEFINITION A.1. (Di Crescenzo 1999). Let X and Y be two non-negative random variables with $X \leq_{st} Y$ and $E[X] < E[Y] < \infty$. Then, $Z \equiv \Psi(X,Y)$ is a random variable with probability density function

$$f_Z(x) = \frac{F_X(x) - F_Y(x)}{E[Y] - E[X]}, x \geq 0,$$

where $F_X(\cdot)$ and $F_Y(\cdot)$ are the cumulative distribution functions of X and Y, respectively. Di Crescenzo (1999) shows that $f_Z(\cdot)$ is a probability density function.

LEMMA A.2. *(Theorem 4.1 of Di Crescenzo 1999) Let X and Y be two non-negative random variables satisfying $X \leq_{st} Y$ and $E[X] < E[Y] < \infty$, and let $Z = \Psi(X,Y)$. Let also g be a measurable and differentiable function such that $E[g(X)]$ and $E[g(Y)]$ are finite, and let its derivative g' be measurable and Riemann-integrable on the interval $[x,y]$ for all $0 \leq x \leq y$. Then, $E\big[g'(Z)\big]$ is finite and*

$$E[g(Y)] - E[g(X)] = E[g'(Z)]\big(E[Y] - E[X]\big). \tag{EC.2}$$

Lemma A.2 presents a probabilistic analogue of the mean value theorem, where Z is a random variable that can be considered as the "mean value" of X and Y. However, unlike for the (deterministic) mean value theorem, Z does not change with the function g, and $Z = \Psi(X,Y)$ is not necessarily ordered (in some stochastic sense) between X and Y. For example, when X and Y are exponential random variables with distinct rates, $Z =_{st} X + Y$ (see Example 3.1 in Di Crescenzo 1999).

A.2. Proofs of results and supplemental material for Sections 3 and 4

Proof of equivalence of Equations (1) ***and*** (2)*:* The long-run average cost in (1) can be written as

$$C_\pi = \sum_{i=1}^{K} \lim_{t\to\infty} \frac{\sum_{k=1}^{n_i(t)} C_i(V_{i,k}^{\pi,x_0})}{n_i(t)} \lim_{t\to\infty} \frac{n_i(t)}{t} = \sum_{i=1}^{K} \lambda p_i \lim_{n\to\infty} \frac{\sum_{k=1}^{n} C_i(V_{i,k}^{\pi,x_0})}{n}, \tag{EC.3}$$

which follows from the fact that $\{n_i(t), t \geq 0\}$ is a Poisson process with rate λp_i for $i \in \{1,2,\ldots,K\}$. In the following, we will prove that for $i \in \{1,2,\ldots,K\}$ when $E\Big[\big|C_i(W_i^\pi)\big|\Big]$ is finite,

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} C_i(V_{i,k}^{\pi,x_0})}{n} = E\left[C_i(W_i^\pi)\right], \tag{EC.4}$$

which shows that (EC.3) (and hence (1)) is equivalent to (2).

In the remainder of this proof, we drop the superscripts π and x_0 for notational convenience, and let T_{ik}, S_{ik} and D_{ik} be the arrival time, service time and departure time of the kth type i customer, respectively, under policy π and initial state x_0. Then, $V_{ik} = D_{ik} - T_{ik} - S_{ik}$ is the queue-waiting time for this customer. Note that $\{V_{ik}, k = 1,2,\ldots\}$ for each $i \in \{1,2,\ldots,K\}$ is a delayed regenerative process with nth regeneration happening at $N_{i,n}$ for $n = 0,1,2,\ldots$, where $N_{i,0} = 1$, and $N_{i,n} = \min\{k : k > N_{i,n-1}, V_{ik} = 0\}$. Note also that for each i, $\{C_i(V_{ik}), k = 1,2,\ldots\}$ is a regenerative process with the same regeneration epoches as $\{V_{ik}\}$. Then, by Theorem 13 of Chapter 2 and last paragraph of page 93 in Wolff (1989), (EC.4) holds for $i \in \{1,2,\ldots,K\}$ if $\sum_{k=1}^{N_{i,1}-1} |C_i(V_{ik})| < \infty$ with probability one, $E[N_{i,2} - N_{i,1}] < \infty$, and $E\left[\sum_{k=N_{i,1}}^{N_{i,2}-1} |C_i(V_{ik})|\right] < \infty$. We next complete the proof by showing that these three conditions hold.

When $\rho < 1$, the system will return to the empty state within finite time with probability one and also the expected time for this return is finite (see, e.g., Theorem 7.11 in Kulkarni (2009)). This implies that $N_{i,1} < \infty$ with probability one, $N_{i,2} - N_{i,1} < \infty$ with probability one, $V_{i,k} < \infty$ for any i and k with probability one and $E[N_{i,2} - N_{i,1}] < \infty$. At last, by Theorem B.5 (i) in El-Taha and Stidham Jr (1999), $E\left[\sum_{k=N_{i,1}}^{N_{i,2}-1} |C_i(V_{ik})|\right] = E[|C_i(W_i)|]\,E[N_{i,2} - N_{i,1}]$ is finite under the assumption that $E[|C_i(W_i)|]$ is finite. □

Proof of Proposition 1*:* We consider a policy $\gamma^* \in \Pi^F$ such that $\gamma^*(j) = 1$ for $j \in \{1,2,\ldots,K\} \setminus \{i\}$ and $\gamma^*(i) = 2$. Then, we can conclude that $W_i^\gamma \leq_{st} W_i^{\gamma^*}$ for any policy $\gamma \in \Pi^F$ by a similar interchange argument as in the proof of Lemma A.1 (b). Then, for $1 \leq \ell < \infty$, by Theorem 1.A.3(a) of Shaked and Shanthikumar (2007), we have $E\left[(W_i^\gamma)^\ell\right] \leq E\left[(W_i^{\gamma*})^\ell\right]$ for any policy $\gamma \in \Pi^F$.

Note that Assumption 1 holds for type i for which $C_i(t)$ is in the polynomial form given in Proposition 2 under a policy $\gamma \in \Pi^F$ if $E\left[(W_i^{\gamma*})^\ell\right]$ is finite for all $\ell = 1,2,\ldots,J_i$. Note also that

$$E\left[\left(W_i^{\gamma^*}\right)^\ell\right] = (-1)^\ell \frac{d^\ell \widetilde{W}_i^{\gamma^*}(s)}{ds^\ell}\Big|_{s=0}. \tag{EC.5}$$

We next define new notation to provide an expression for $\widetilde{W}_i^{\gamma*}(s)$. Let $\tilde{B}(s)$ denote the LST of the length of a busy period starting from an empty and idle system, and let $\tilde{B}_{\neq i}(s)$ denote the LST of $T_{B_{\neq i}}$, the length of a busy period where only type $j \in \{1,2,\ldots,K\} \setminus \{i\}$ customers arrive and are served according

to FCFS. Let W^F denote the steady-state waiting time under FCFS and $\widetilde{W}^F(\cdot)$ be its LST. Finally, let $\tilde{S}(x)=\sum_{j=1}^K p_j\tilde{S}_j(x)$. From Equation (3.8) and (3.10) of Miller (1960), we have

$$\widetilde{W}_i^{\gamma^*}(s)=\widetilde{W}^F\big(\lambda(1-p_i)(1-\tilde{B}_{\neq i}(s))+s\big),$$

where $\widetilde{W}^F(s)=(1-\rho)s\Big/\Big[s-\lambda\Big(1-\tilde{S}(s)\Big)\Big]$, and $\tilde{B}_{\neq i}(s)$ is the unique solution to $\tilde{B}_{\neq i}(s)=\tilde{S}_{\neq i}\big(s+\lambda(1-p_i)(1-\tilde{B}_{\neq i}(s))\big)$ for $s>0$ and $\lim_{s\to\infty}\tilde{B}_{\neq i}(s)=0$, and $\tilde{S}_{\neq i}(s)=\big(\tilde{S}(s)-p_i\tilde{S}_i(s)\big)/(1-p_i)$. Then, using Faa di Bruno's formula (see, e.g., Theorem 2 of Roman (1980)), (EC.5) is finite if $\frac{d^n\widetilde{W}^F(s)}{ds^n}|_{s=0}$ and $\frac{d^n\tilde{B}_{\neq i}(s)}{ds^n}|_{s=0}$ are finite for all $n\le\ell$, i.e., if the nth moment of W^F and $T_{B_{\neq i}}$ are finite.

When $\rho<1$, we can obtain the nth moment of W^F as (see, e.g., page 238 in Gross et al. (2008))

$$E\left[\left(W^F\right)^n\right]=\frac{\lambda}{1-\rho}\sum_{\ell=1}^{n}\binom{n}{\ell}E\left[\left(W^F\right)^{n-\ell}\right]\frac{E\left[S^{\ell+1}\right]}{\ell+1},$$

where $E\left[S^{\ell+1}\right]$ is the $(\ell+1)$st moment of service time of a randomly picked customer. Hence, $E\left[(W^F)^n\right]$ is finite if $\rho<1$ and the first $n+1$ moments of service times of all customers are finite. Besides, from Theorem 1 of Ghahramani and Wolff (1989), the nth moment of the busy period is finite if and only if the nth moment of the service times is finite. Thus, $E\left[\left(W_i^{\gamma^*}\right)^\ell\right]$ is finite if $\rho<1$ and the first $(\ell+1)$ moments of service times are finite. □

Proof of Proposition 2: We only prove part (a) here because the proofs of parts (b) and (c) are very similar. From Equation (2), we have

$$\begin{aligned}C_{\pi_k}-C_{\bar{\pi}_k}&=\lambda\sum_{j=1}^{K}p_j\Big(E\big[C_j(W_j^{\pi_k})\big]-E\big[C_j(W_j^{\bar{\pi}_k})\big]\Big)\\&=\lambda p_k\Big(E\big[C_k(W_k^{\pi_k})\big]-E\big[C_k(W^{\bar{\pi}_k})\big]\Big)+\lambda p_{k+1}\Big(E\big[C_{k+1}(W_{k+1}^{\pi_k})\big]-E\big[C_{k+1}(W^{\bar{\pi}_k})\big]\Big),\end{aligned}$$

where the last equation follows from Lemma A.1 (a). Then, $C_{\pi_k}\le C_{\bar{\pi}_k}$ if and only if

$$p_{k+1}\Big(E\big[C_{k+1}(W_{k+1}^{\pi_k})\big]-E\big[C_{k+1}(W^{\bar{\pi}_k})\big]\Big)\le p_k\Big(E\big[C_k(W^{\bar{\pi}_k})\big]-E\big[C_k(W_k^{\pi_k})\big]\Big).\tag{EC.6}$$

According to the work conservation law (see, e.g., Kleinrock (1965)), we have

$$\sum_{j=1}^{K}p_jE\big[W_j^{\pi_k}\big]/\mu_j=\sum_{j=1}^{K}p_jE\big[W_j^{\bar{\pi}_k}\big]/\mu_j,$$

and by Lemma A.1 (a), we have $E\big[W_j^{\pi_k}\big]=E\big[W_j^{\bar{\pi}_k}\big]$ for $j\in\{1,2,\ldots,K\}\setminus\{k,k+1\}$. Hence,

$$p_{k+1}\Big(E\big[W_{k+1}^{\pi_k}\big]-E\big[W^{\bar{\pi}_k}\big]\Big)/\mu_{k+1}=p_k\Big(E\big[W^{\bar{\pi}_k}\big]-E\big[W_k^{\pi_k}\big]\Big)/\mu_k,\tag{EC.7}$$

which is positive by Lemma A.1 (c). Dividing (EC.6) by (EC.7) completes the proof. □

REMARK A.1. For $i\in\{k,k+1\}$, $\gamma_1,\gamma_2\in\{\pi_k,\pi_{k+1},\bar{\pi}_k\}$, and $\gamma_1\neq\gamma_2$, let $U_i^{\gamma_1,\gamma_2}\equiv\Psi\left(W_i^{\gamma_1},W_i^{\gamma_2}\right)$, where $\Psi(\cdot,\cdot)$ is defined in Definition A.1. Note that $U_i^{\gamma_1,\gamma_2}$ is well defined for $i\in\{k,k+1\}$ and $\gamma_1,\gamma_2\in\{\pi_k,\pi_{k+1},\bar{\pi}_k\}$ because $E\big[W_i^{\pi_i}\big]<E\big[W^{\bar{\pi}_k}\big]<E\big[W_i^{\pi_{2k+1-i}}\big]$ from Lemma A.1 (c), and due to the stochastic ordering provided in Lemma A.1 (b). Then, from Lemma A.2, we have $\delta_i^{\gamma_1,\gamma_2}=E\left[C_i'(U_i^{\gamma_1,\gamma_2})\right]$ for $i\in\{k,k+1\}$ and $\gamma_1,\gamma_2\in\{\pi_k,\pi_{k+1},\bar{\pi}_k\}$.

In an immediate corollary to Proposition 2, we provide necessary and sufficient conditions for the optimality of π_k, π_{k+1}, and $\bar{\pi}_k$ within the set of these three policies, for which we only need to calculate the values of $\delta_i^\gamma\equiv\delta_i^{\gamma,\bar{\pi}_k}$ for $i\in\{k,k+1\}$ and $\gamma\in\{\pi_k,\pi_{k+1}\}$.

COROLLARY A.1. *Let $I(\bar{\pi}_k)\equiv 0$ and $I(\pi_i)\equiv\left(\delta_i^{\pi_i}\mu_i-\delta_j^{\pi_i}\mu_j\right)/\left(\bar{\rho}_{k-1}-\rho_i\right)$ for $i,j\in\{k,k+1\}$ and $j\neq i$. Then, the policy with the largest [smallest] value of $I(\gamma)$ for $\gamma\in\{\pi_k,\pi_{k+1},\bar{\pi}_k\}$ has the lowest [highest] long-run average cost among these three policies.*

Proof of Corollary A.1: For $i \in \{k, k+1\}$, $I(\bar{\pi}_k) \geq I(\pi_i) \Leftrightarrow \delta_i^{\pi_i}\mu_i \leq \delta_j^{\pi_i}\mu_j \Leftrightarrow C_{\bar{\pi}_k} \leq C_{\pi_i}$, which follows from Proposition 2 (a) and (b). Furthermore, for $i, j \in \{k, k+1\}$ and $i \neq j$, we get

$$\delta_i^{\pi_k,\pi_{k+1}} = \delta_i^{\pi_j}\left(\frac{E\left[W_i^{\pi_j}\right] - E\left[W^{\bar{\pi}_k}\right]}{E\left[W_i^{\pi_j}\right] - E\left[W_i^{\pi_i}\right]}\right) + \delta_i^{\pi_i}\left(\frac{E\left[W^{\bar{\pi}_k}\right] - E\left[W_i^{\pi_i}\right]}{E\left[W_i^{\pi_j}\right] - E\left[W_i^{\pi_i}\right]}\right) = \frac{\delta_i^{\pi_j}\left(\bar{\rho}_{k-1} - \rho_i\right) + \delta_i^{\pi_i}\left(\bar{\rho}_{k-1} - \rho_j\right)}{2\bar{\rho}_{k-1} - \rho_k - \rho_{k+1}},$$

which follows from Lemma A.1 (c). Then, by Proposition 2 (c)

$$\begin{aligned} I(\pi_k) \geq I(\pi_{k+1}) &\Leftrightarrow \mu_k\left[\delta_k^{\pi_k}\left(\bar{\rho}_{k-1} - \rho_{k+1}\right) + \delta_k^{\pi_{k+1},\bar{\pi}_k}\left(\bar{\rho}_{k-1} - \rho_k\right)\right] \geq \mu_{k+1}\left[\delta_{k+1}^{\pi_k}\left(\bar{\rho}_{k-1} - \rho_{k+1}\right) + \delta_{k+1}^{\pi_{k+1},\bar{\pi}_k}\left(\bar{\rho}_{k-1} - \rho_k\right)\right] \\ &\Leftrightarrow \delta_k^{\pi_k,\pi_{k+1}}\mu_k \geq \delta_{k+1}^{\pi_k,\pi_{k+1}}\mu_{k+1} \Leftrightarrow C_{\pi_k} \leq C_{\pi_{k+1}}. \end{aligned}$$

□

A.3. Proofs of results and supplemental material for Sections 5 and 6

Proof of Proposition 3: Using Equation (69) of Takács (1964), we obtain expressions for $E\left[(W^{\bar{\pi}_k})^2\right], E\left[(W_i^{\pi_i})^2\right]$, and $E\left[(W_j^{\pi_i})^2\right]$ for $i, j \in \{k, k+1\}$ and $i \neq j$, which lead to the following when combined with (4) and Lemma A.1 (c):

$$M_i^{\pi_i} = \frac{1}{\bar{\rho}_{k-1}}\left[\frac{2\tau^{(3)}}{3\tau^{(2)}} + \frac{\lambda\sum_{\ell=1}^{k+1} p_\ell\tau_\ell^{(2)}}{\bar{\rho}_{k+1}} + \frac{\lambda\left(\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)} + p_i\tau_i^{(2)}\right)}{\bar{\rho}_{k-1} - \rho_i} + \frac{\lambda\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)}}{\bar{\rho}_{k-1}} + \tau_j^{(2)}\mu_j\right], \tag{EC.8}$$

$$M_j^{\pi_i} = \frac{1}{\bar{\rho}_{k-1}}\left[\left(\frac{2\bar{\rho}_{k-1} - \rho_i}{\bar{\rho}_{k-1} - \rho_i}\right)\left(\frac{2\tau^{(3)}}{3\tau^{(2)}} + \frac{\lambda\sum_{\ell=1}^{k+1} p_\ell\tau_\ell^{(2)}}{\bar{\rho}_{k+1}} + \frac{\lambda\left(\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)} + p_i\tau_i^{(2)}\right)}{\bar{\rho}_{k-1} - \rho_i}\right) + \frac{\lambda\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)}}{\bar{\rho}_{k-1}} + \tau_i^{(2)}\mu_i\right]. \tag{EC.9}$$

We next show that $\delta_i^{\pi_i} < \delta_i^{\pi_j}$ when $c_i^{(2)} > 0$ for $i \in \{k, k+1\}$ and $j = 2k+1-i$ by showing that $M_i^{\pi_i} < M_i^{\pi_j}$. By switching the indices of i and j in (EC.9), and subtracting (EC.8), we have

$$\begin{aligned} M_i^{\pi_j} - M_i^{\pi_i} = \frac{2\tau^{(3)}}{3\tau^{(2)}(\bar{\rho}_{k-1} - \rho_j)} &+ \frac{\lambda\rho_j\left(\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)} + p_i\tau_i^{(2)}\right)}{\bar{\rho}_{k-1}\bar{\rho}_{k+1}}\left(\frac{1}{\bar{\rho}_{k-1} - \rho_i} + \frac{1}{\bar{\rho}_{k-1} - \rho_j}\right) \\ &+ \frac{\lambda p_j\tau_j^{(2)}}{(\bar{\rho}_{k-1} - \rho_j)\bar{\rho}_{k+1}} + \frac{\lambda\left(\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)} + p_j\tau_j^{(2)}\right)}{\bar{\rho}_{k-1} - \rho_j}\left(\frac{1}{\bar{\rho}_{k-1}} + \frac{1}{\bar{\rho}_{k-1} - \rho_j}\right), \end{aligned} \tag{EC.10}$$

which is positive because $0 < \rho_j < \bar{\rho}_{k-1}$, $0 < \rho_i < \bar{\rho}_{k-1}$, $\bar{\rho}_{k+1} > 0$, and all moments of service times are positive.

Finally, when both cost functions for type k and $k+1$ are quadratic, if $\delta_i^{\pi_i}\mu_i \geq \delta_j^{\pi_i}\mu_j$, for some $i \in \{k, k+1\}$ and $j = 2k+1-i$ (and thus $\delta_i^{\pi_j}\mu_i > \delta_i^{\pi_i}\mu_i \geq \delta_j^{\pi_i}\mu_j > \delta_j^{\pi_j}\mu_j$), then $I(\pi_i) > I(\bar{\pi}_k) > I(\pi_j)$, and hence π_i is the best and π_j is the worst according to Corollary A.1. On the other hand, if $\delta_k^{\pi_k}\mu_k \leq \delta_{k+1}^{\pi_k}\mu_{k+1}$ and $\delta_{k+1}^{\pi_{k+1}}\mu_{k+1} \leq \delta_k^{\pi_{k+1}}\mu_k$, then $I(\pi_k) \leq I(\bar{\pi}_k)$ and $I(\pi_{k+1}) \leq I(\bar{\pi}_k)$, and hence $\bar{\pi}_k$ is the best by Corollary A.1. □

Proof of Corollary 1: Under Assumption 2, we have $\delta_i^{\pi_i}\mu_i \geq \delta_j^{\pi_i}\mu_j$ if and only if $\mu_i c_i^{(2)} M_i^{\pi_i} \geq \mu_j c_j^{(2)} M_j^{\pi_i}$ for $i \in \{k, k+1\}$ and $j = 2k+1-i$. Then, the expressions for $R^{\pi_{k+1}}$ and R^{π_k} and the characterization of the best/worst policy follow directly from Proposition 3. From (EC.10), we have $M_{k+1}^{\pi_{k+1}} < M_{k+1}^{\pi_k}$ and $M_k^{\pi_{k+1}} > M_k^{\pi_k}$, and thus, $R^{\pi_{k+1}} = \dfrac{M_{k+1}^{\pi_{k+1}}}{M_k^{\pi_{k+1}}} < \dfrac{M_{k+1}^{\pi_k}}{M_k^{\pi_k}} = R^{\pi_k}$.

When $\mu_k = \mu_{k+1}$ and $\tau_k^{(2)} = \tau_{k+1}^{(2)}$, for $i \in \{k, k+1\}$ and $j = 2k+1-i$, (EC.8) and (EC.9) yield

$$M_j^{\pi_i} - M_i^{\pi_i} = \frac{1}{\bar{\rho}_{k-1} - \rho_i}\left(\frac{2\tau^{(3)}}{3\tau^{(2)}} + \frac{\lambda\sum_{j=1}^{k+1} p_j\tau_j^{(2)}}{\bar{\rho}_{k+1}} + \frac{\lambda\left(\sum_{j=1}^{k-1} p_j\tau_j^{(2)} + p_i\tau_i^{(2)}\right)}{\bar{\rho}_{k-1} - \rho_i}\right) > 0.$$

Hence, $R^{\pi_{k+1}} < 1 < R^{\pi_k}$. □

Proof of Proposition 4: We first show that for $i \in \{1, 2\}$, $\dfrac{\partial}{\partial\lambda}\left(\dfrac{M_i^{\pi_i}}{M_{3-i}^{\pi_i}}\right) < 0$ if

$$\frac{\tau_{3-i}^{(2)}\mu_{3-i}}{\tau_i^{(2)}\mu_i} \geq \frac{1 - 2\rho_i}{2(1 - \rho_i)}. \tag{EC.11}$$

We define $G_i(\lambda)$ for $i\in\{1,2\}$ as $G_i(\lambda)=\dfrac{2\tau^{(3)}}{3\tau^{(2)}}+\dfrac{\lambda\tau^{(2)}}{1-\rho}+\dfrac{\lambda p_i\tau_i^{(2)}}{1-\rho_i}$. Then, (EC.8) and (EC.9) reduce to

$$M_i^{\pi_i}=G_i(\lambda)+\tau_{3-i}^{(2)}\mu_{3-i} \text{ and } M_{3-i}^{\pi_i}=\left(\frac{2-\rho_i}{1-\rho_i}\right)G_i(\lambda)+\tau_i^{(2)}\mu_i, \text{ for } i\in\{1,2\}.$$

Then, we have

$$\frac{\partial}{\partial\lambda}\left(\frac{M_i^{\pi_i}}{M_{3-i}^{\pi_i}}\right)=\frac{G_i'(\lambda)\left(\tau_i^{(2)}\mu_i-\left(\frac{2-\rho_i}{1-\rho_i}\right)\tau_{3-i}^{(2)}\mu_{3-i}\right)-\left(G_i(\lambda)+\tau_{3-i}^{(2)}\mu_{3-i}\right)\frac{p_i}{\mu_i(1-\rho_i)^2}G_i(\lambda)}{\left(\left(\frac{2-\rho_i}{1-\rho_i}\right)G_i(\lambda)+\tau_i^{(2)}\mu_i\right)^2}<0$$

if and only if

$$\tau_i^{(2)}\mu_i \;-\; \left(\frac{2-\rho_i}{1-\rho_i}\right)\tau_{3-i}^{(2)}\mu_{3-i} \;<\; \frac{\frac{p_i}{\mu_i(1-\rho_i)^2}\left(\frac{2\tau^{(3)}}{3\tau^{(2)}}+\frac{\lambda\tau^{(2)}}{1-\rho}+\frac{\lambda p_i\tau_i^{(2)}}{1-\rho_i}\right)\left(\frac{2\tau^{(3)}}{3\tau^{(2)}}+\frac{\lambda\tau^{(2)}}{1-\rho}+\frac{\lambda p_i\tau_i^{(2)}}{1-\rho_i}+\tau_{3-i}^{(2)}\mu_{3-i}\right)}{\frac{\tau^{(2)}}{(1-\rho)^2}+\frac{p_i\tau_i^{(2)}}{(1-\rho_i)^2}}, \tag{EC.12}$$

because for $i\in\{1,2\}$, $G_i'(\lambda)=\dfrac{\tau^{(2)}}{(1-\rho)^2}+\dfrac{p_i\tau_i^{(2)}}{(1-\rho_i)^2}>0$. Note that the right-hand side of (EC.12) is greater than

$$\begin{aligned}\frac{\frac{p_i}{\mu_i(1-\rho_i)^2}\left(\frac{\lambda\tau^{(2)}}{1-\rho}+\frac{\lambda p_i\tau_i^{(2)}}{1-\rho_i}\right)\left(\frac{\lambda\tau^{(2)}}{1-\rho}+\frac{\lambda p_i\tau_i^{(2)}}{1-\rho_i}+\tau_{3-i}^{(2)}\mu_{3-i}\right)}{\frac{\tau^{(2)}}{(1-\rho)^2}+\left(\frac{1-\rho_i}{1-\rho}\right)\frac{p_i\tau_i^{(2)}}{(1-\rho_i)^2}}&=\tau_i^{(2)}\mu_i\left(\frac{\rho_i}{1-\rho_i}\right)^2\left(1+\frac{1-\rho}{1-\rho_i}\right)+\tau_{3-i}^{(2)}\mu_{3-i}\left(\frac{\rho_i}{1-\rho_i}\right)\\&>\tau_i^{(2)}\mu_i\left(\frac{\rho_i}{1-\rho_i}\right)^2+\tau_{3-i}^{(2)}\mu_{3-i}\left(\frac{\rho_i}{1-\rho_i}\right).\end{aligned}$$

Thus, a sufficient condition for (EC.12) to hold is

$$\tau_i^{(2)}\mu_i-\left(\frac{2-\rho_i}{1-\rho_i}\right)\tau_{3-i}^{(2)}\mu_{3-i}\le\tau_i^{(2)}\mu_i\left(\frac{\rho_i}{1-\rho_i}\right)^2+\tau_{3-i}^{(2)}\mu_{3-i}\left(\frac{\rho_i}{1-\rho_i}\right),$$

which reduces to (EC.11).

Now note that $R^{\pi_1}=M_2^{\pi_1}/M_1^{\pi_1}$ increases in λ if and only if $M_1^{\pi_1}/M_2^{\pi_1}$ decreases in λ. Then, by letting $i=1$ in (EC.11) we obtain part (a). Similarly, by letting $i=2$ in (EC.11), we obtain part (b).

Finally, to prove part (c), note that as $\lambda\to\mu$, we have $G_i(\lambda)\to\infty$, and hence,

$$\lim_{\lambda\to\mu}\frac{M_i^{\pi_i}}{M_{3-i}^{\pi_i}}=\lim_{\lambda\to\mu}\frac{G_i(\lambda)+\tau_{3-i}^{(2)}\mu_{3-i}}{\left(\frac{2-\rho_i}{1-\rho_i}\right)G_i(\lambda)+\tau_i^{(2)}\mu_i}=\lim_{\lambda\to\mu}\frac{G_i(\lambda)}{\left(\frac{2-\rho_i}{1-\rho_i}\right)G_i(\lambda)}=\lim_{\lambda\to\mu}\frac{1-\rho_i}{2-\rho_i}=\lim_{\lambda\to\mu}\frac{p_{3-i}/\mu_{3-i}}{p_i/\mu_i+2p_{3-i}/\mu_{3-i}}.$$

Then, letting $i=1$ and $i=2$ provides the limits for R^{π_1} and R^{π_2} as $\lambda\to\mu$. □

Proof of Proposition 5: Since the first three moments of service times do not depend on type, we drop the subscript from μ_i, $\tau_i^{(2)}$, and $\tau_i^{(3)}$. Then, we have

$$R^{\pi_1}=\frac{\left(\frac{2-\rho_1}{1-\rho_1}\right)M+\frac{\tau^{(2)}\mu}{(1-\rho_1)^2}}{M+\frac{\tau^{(2)}\mu}{(1-\rho_1)}},\;\; R^{\pi_2}=\frac{M+\frac{\tau^{(2)}\mu}{(1-\rho_2)}}{\left(\frac{2-\rho_2}{1-\rho_2}\right)M+\frac{\tau^{(2)}\mu}{(1-\rho_2)^2}},$$

where $M\equiv\dfrac{2\tau^{(3)}}{3\tau^{(2)}}+\dfrac{\lambda\tau^{(2)}}{1-\rho}$, which is a positive constant independent of p_i for $i\in\{1,2\}$. Then, we have

$$\frac{\partial R^{\pi_1}}{\partial p_1}=\frac{\frac{M^2}{\mu}+\frac{\rho_1\tau^{(2)}}{(1-\rho_1)}M}{\frac{(1-\rho_1)^2}{\lambda}\left[M+\frac{\tau^{(2)}\mu}{(1-\rho_1)}\right]^2}>0 \text{ and } \frac{\partial R^{\pi_2}}{\partial p_2}=\frac{-\frac{M^2}{\mu}-\frac{\rho_2\tau^{(2)}}{(1-\rho_2)}M}{\frac{(1-\rho_2)^2}{\lambda}\left[\left(\frac{2-\rho_2}{1-\rho_2}\right)M+\frac{\tau^{(2)}\mu}{(1-\rho_2)^2}\right]^2}<0.$$

□

Proof of Proposition 6: When service times are exponential, $\tau_i^{(3)} = 6/\mu_i^3$ and $\tau_i^{(2)} = 2/\mu_i^2$; and when service times are deterministic, $\tau_i^{(3)} = 1/\mu_i^3$ and $\tau_i^{(2)} = 1/\mu_i^2$ for $i \in \{1,2\}$, and hence we have,

$$R_{exp}^{\pi_1} = \frac{\left(\frac{2-\rho_1}{1-\rho_1}\right) N_{exp}^{(1)} + \frac{1}{\mu_1}}{N_{exp}^{(1)} + \frac{1}{\mu_2}},\ R_{exp}^{\pi_2} = \frac{N_{exp}^{(2)} + \frac{1}{\mu_1}}{\left(\frac{2-\rho_2}{1-\rho_2}\right) N_{exp}^{(2)} + \frac{1}{\mu_2}};\quad R_{det}^{\pi_1} = \frac{\left(\frac{2-\rho_1}{1-\rho_1}\right) N_{det}^{(1)} + \frac{1}{\mu_1}}{N_{det}^{(1)} + \frac{1}{\mu_2}},\ R_{det}^{\pi_2} = \frac{N_{det}^{(2)} + \frac{1}{\mu_1}}{\left(\frac{2-\rho_2}{1-\rho_2}\right) N_{det}^{(2)} + \frac{1}{\mu_2}}.$$

Here, for $i \in \{1,2\}$,

$$N_{exp}^{(i)} \equiv \frac{p_1/\mu_1^3 + p_2/\mu_2^3}{p_1/\mu_1^2 + p_2/\mu_2^2} + \frac{\lambda\left(p_1/\mu_1^2 + p_2/\mu_2^2\right)}{1-\rho} + \frac{\lambda p_i/\mu_i^2}{1-\rho_i},\quad N_{det}^{(i)} \equiv \frac{2\left(p_1/\mu_1^3 + p_2/\mu_2^3\right)}{3\left(p_1/\mu_1^2 + p_2/\mu_2^2\right)} + \frac{\lambda\left(p_1/\mu_1^2 + p_2/\mu_2^2\right)}{1-\rho} + \frac{\lambda p_i/\mu_i^2}{1-\rho_i},$$

where $N_{exp}^{(i)} > N_{det}^{(i)}$. Taking the difference of $R_{exp}^{\pi_1}$ and $R_{det}^{\pi_1}$, we have

$$R_{exp}^{\pi_1} - R_{det}^{\pi_1} = \frac{\left(\frac{1}{\mu_2}\left(\frac{2-\rho_1}{1-\rho_1}\right) - \frac{1}{\mu_1}\right)\left(N_{exp}^{(1)} - N_{det}^{(1)}\right)}{\left(N_{det}^{(1)} + \frac{1}{\mu_2}\right)\left(N_{exp}^{(1)} + \frac{1}{\mu_2}\right)}.$$

Hence, $R_{exp}^{\pi_1} \geq R_{det}^{\pi_1}$ if and only if $\mu_2/\mu_1 \leq (2-\rho_1)/(1-\rho_1)$, which proves part (a). Part (b) can be proved similarly by obtaining $R_{exp}^{\pi_2} - R_{det}^{\pi_2}$. □

We first provide Lemma A.3 to facilitate the proof of Proposition 7.

LEMMA A.3. *Suppose that λ, μ_i, $\tau_i^{(2)}$, and $\tau_i^{(3)}$ are finite and $\mu_i = \mu > \lambda$ for all $i = 1,2,\ldots,K$. Then,*
(a) $\pi_k^* = \pi_i$ if and only if $f_i(\tau_i^{(2)}, \tau_j^{(2)}) > 0$ for $i \in \{k, k+1\}$ and $j = 2k+1-i$, where for $x_1, x_2 > 0$,

$$f_i(x_1,x_2) \equiv \left(\bar\rho_{k-1}\bar\rho_{k+1} - (\rho_k + \rho_{k+1})(\bar\rho_{k-1} - \rho_i)\right)x_2 - \left(\rho_i(\rho_k + \rho_{k+1}) + \frac{\bar\rho_{k-1}^2\bar\rho_{k+1}}{\bar\rho_{k-1} - \rho_i}\right)x_1$$
$$- \bar\rho_{k-1}\bar\rho_{k+1}\left(\frac{2\tau^{(3)}}{3\mu\left(\sum_{\ell=1,\ell\notin\{k,k+1\}}^K p_\ell\tau_\ell^{(2)} + p_i x_1 + p_j x_2\right)}\right) - \bar\rho_{k-1}\rho\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)}\left(1 + \frac{\bar\rho_{k+1}}{\bar\rho_{k-1} - \rho_i}\right). \quad \text{(EC.13)}$$

(b) For $i \in \{k, k+1\}$, $f_i(x_1,x_2) < 0$ for any $x_1, x_2 > 0$ if

$$\rho \geq \frac{1}{\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_{2k+1-i}}}, \quad \text{(EC.14)}$$

which is feasible for $\rho < 1$ when $p_k + p_{k+1} > \sum_{\ell=k+2}^K p_\ell$ and $p_{2k+1-i} > \left(\sum_{\ell=k+2}^K p_\ell\right)^2/(p_k + p_{k+1})$.
(c) For $i \in \{k, k+1\}$, if (EC.14) *holds in the opposite direction, then for any fixed $x_1 > 0$, there exists $h_i(x_1) \in (x_1, \infty)$ such that $f_i(x_1, h_i(x_1)) = 0$, and $f_i(x_1,x_2) > 0$ if and only if $x_2 > h_i(x_1)$. Furthermore, if $p_i \leq p_{2k+1-i}$, then $h_i(x_1)$ strictly increases in x_1.*

Proof of Lemma A.3: (a) By Corollary 1, $\pi_k^* = \pi_i$ if and only if $M_i^{\pi_i} > M_j^{\pi_i}$ for $i \in \{k, k+1\}$ and $j = 2k+1-i$. Taking the difference of (EC.8) and (EC.9), and using the fact that $\mu_i = \mu$ for all $i = 1,2,\ldots,K$, we have

$$M_i^{\pi_i} - M_j^{\pi_i} = \frac{\mu(\tau_j^{(2)} - \tau_i^{(2)})}{\bar\rho_{k-1}} - \frac{1}{\bar\rho_{k-1} - \rho_i}\left(\frac{2\tau^{(3)}}{3\tau^{(2)}} + \frac{\lambda\sum_{\ell=1}^{k+1} p_\ell\tau_\ell^{(2)}}{\bar\rho_{k+1}} + \frac{\lambda\left(\sum_{\ell=1}^{k-1} p_\ell\tau_\ell^{(2)} + p_i\tau_i^{(2)}\right)}{\bar\rho_{k-1} - \rho_i}\right) > 0$$

if and only if $f_i(\tau_i^{(2)}, \tau_j^{(2)}) > 0$ for $i \in \{k, k+1\}$ and $j = 2k+1-i$.

(b) By (EC.13), $f_i(x_1,x_2) < 0$ for any $x_1, x_2 > 0$ if $\bar\rho_{k-1}\bar\rho_{k+1} - (\rho_k + \rho_{k+1})(\bar\rho_{k-1} - \rho_i) \leq 0$, which reduces to (EC.14). For $i \in \{k, k+1\}$, (EC.14) could hold for $\rho < 1$ only if its right-hand side is less than one, i.e.,

$$\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_{2k+1-i}} > 1 \Leftrightarrow \frac{p_{2k+1-i}}{p_k + p_{k+1}} > \left(\frac{\sum_{\ell=k+2}^K p_\ell}{p_k + p_{k+1}}\right)^2,$$

which holds only if we have $p_k + p_{k+1} > \sum_{j=k+2}^K p_j$.

(c) If (EC.14) holds in the opposite direction for $i \in \{k, k+1\}$, it is directly observed from (EC.13) that for fixed x_1, $f_i(x_1, x_2)$ increases in x_2. Also, for any fixed $x_1 > 0$, from (EC.13) we obtain $f_i(x_1, x_2) \to \infty$ as $x_2 \to \infty$ and

$$f_i(x_1, x_1) < \left(\bar{\rho}_{k-1}\bar{\rho}_{k+1} - \frac{\bar{\rho}_{k-1}^2\bar{\rho}_{k+1}}{\bar{\rho}_{k-1} - \rho_i}\right) x_1 = -\left(\frac{\bar{\rho}_{k-1}\bar{\rho}_{k+1}\rho_i}{\bar{\rho}_{k-1} - \rho_i}\right) x_1 < 0.$$

Hence, for any fixed $x_1 > 0$, there exists unique $h_i(x_1) \in (x_1, \infty)$ such that $f_i(x_1, h_i(x_1)) = 0$, and for any $x_2 > h_i(x_1)$, $f_i(x_1, x_2) > 0$ and for any $x_2 \le h_i(x_1)$, $f_i(x_1, x_2) \le 0$.

Next, we show that $h_i(x_1)$ increases in x_1 if $p_i \le p_j$ for $i, j \in \{k, k+1\}$ and $i \ne j$. First, define

$$\alpha_i \equiv \big(\bar{\rho}_{k-1}\bar{\rho}_{k+1} - (\rho_k + \rho_{k+1})(\bar{\rho}_{k-1} - \rho_i)\big),\ \beta \equiv \frac{2\bar{\rho}_{k-1}\bar{\rho}_{k+1}\tau^{(3)}}{3\mu},\ \Theta_i \equiv \sum_{\ell=1, \ell \notin \{k,k+1\}}^{K} p_\ell \tau_\ell^{(2)} + p_i x_1,$$

$$D_i \equiv \left(\rho_i(\rho_k + \rho_{k+1}) + \frac{\bar{\rho}_{k-1}^2\bar{\rho}_{k+1}}{\bar{\rho}_{k-1} - \rho_i}\right) x_1 + \bar{\rho}_{k-1}\rho \sum_{\ell=1}^{k-1} p_\ell \tau_\ell^{(2)} \left(1 + \frac{\bar{\rho}_{k+1}}{\bar{\rho}_{k-1} - \rho_i}\right).$$

Then, (EC.13) reduces to $f_i(x_1, x_2) = \alpha_i x_2 - \beta(\Theta_i + p_j x_2)^{-1} - D_i$. Note that when (EC.14) holds in the opposite direction, we have $\alpha_i > 0$. Also, note that setting $f_i(x_1, x_2) = 0$ is equivalent to letting $\alpha_i p_j x_2^2 + (\alpha_i\Theta_i - D_i p_j)x_2 - (D_i\Theta_i + \beta) = 0$. Then, $h_i(x_1)$ is the unique positive root of this quadratic function, i.e.,

$$h_i(x_1) = \frac{-\alpha_i\Theta_i + D_i p_j + \sqrt{\Delta_i}}{2\alpha_i p_j},$$

where $\Delta_i \equiv (D_i p_j - \alpha_i\Theta_i)^2 + 4\alpha_i p_j(D_i\Theta_i + \beta) = (D_i p_j + \alpha_i\Theta_i)^2 + 4\alpha_i p_j\beta > 0$. Note that α_i and β are independent of x_1, $d\Theta_i/dx_1 = p_i > 0$, and

$$\frac{dD_i}{dx_1} = \rho_i(\rho_k + \rho_{k+1}) + \frac{\bar{\rho}_{k-1}^2\bar{\rho}_{k+1}}{\bar{\rho}_{k-1} - \rho_i} > \frac{\bar{\rho}_{k-1}^2\bar{\rho}_{k+1}}{\bar{\rho}_{k-1} - \rho_i} > \bar{\rho}_{k-1}\bar{\rho}_{k+1} > \alpha_i, \quad \frac{d\Delta_i}{dx_1} = 2(D_i p_j + \alpha_i\Theta_i)\left(\frac{dD_i}{dx_1}p_j + \alpha_i p_i\right) > 0.$$

Then,

$$\frac{dh_i(x_1)}{dx_1} = \frac{1}{2\alpha_i p_j}\left(p_j\left(\frac{dD_i}{dx_1}\right) - \alpha_i\left(\frac{d\Theta_i}{dx_1}\right) + \frac{1}{2\sqrt{\Delta_i}}\left(\frac{d\Delta_i}{dx_1}\right)\right) > \frac{1}{2\alpha_i p_j}\left(p_j\left(\frac{dD_i}{dx_1}\right) - \alpha_i p_i\right) > 0,$$

where the last inequality follows because $dD_i/dx_1 > \alpha_i$ and $0 < p_i \le p_j$. □

Proof of Proposition 7: (a) If $\rho \ge \left(\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_k}\right)^{-1}$, then (EC.14) holds for both $i = k$ and $i = k+1$ under the assumption $p_k \le p_{k+1}$. Then, $f_k(\tau_k^{(2)}, \tau_{k+1}^{(2)}) < 0$ and $f_{k+1}(\tau_{k+1}^{(2)}, \tau_k^{(2)}) < 0$ from Lemma A.3 (b), and $\pi_k^* \notin \{\pi_k, \pi_{k+1}\}$ from Lemma A.3 (a). Thus, $\pi_k^* = \bar{\pi}_k$.

For $\rho \ge \left(\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_k}\right)^{-1}$ to hold when $\rho < 1$, we need $p_k + p_{k+1} > \sum_{\ell=k+2}^{K} p_\ell$ and $p_k/(p_k + p_{k+1}) > \left(\sum_{\ell=k+2}^{K} p_\ell/(p_k + p_{k+1})\right)^2$ by Lemma A.3 (b), where the latter inequality is equivalent to

$$\frac{p_k + p_{k+1}}{p_k} < \left(\frac{p_k + p_{k+1}}{\sum_{\ell=k+2}^{K} p_\ell}\right)^2 \Leftrightarrow \frac{p_{k+1}}{p_k} < \left(\frac{p_k + p_{k+1}}{\sum_{\ell=k+2}^{K} p_\ell}\right)^2 - 1.$$

The above inequality holds for $p_k \le p_{k+1}$ only if $p_k + p_{k+1} \ge \sqrt{2}\sum_{\ell=k+2}^{K} p_\ell$.

(b) If $\left(\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_{k+1}}\right)^{-1} \le \rho < \left(\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_k}\right)^{-1}$, (EC.14) holds for $i = k$ (implying $\pi_k^* \ne \pi_k$ by parts (a) and (b) of Lemma A.3) and holds in the opposite direction for $i = k+1$. Now, let $\xi_k \equiv h_{k+1}(\tau_{k+1}^{(2)})$, where $h_{k+1}(\cdot)$ is defined in Lemma A.3 (c) and $\xi_k > \tau_{k+1}^{(2)}$. Then, $f_{k+1}(\tau_{k+1}^{(2)}, \tau_k^{(2)}) > 0$ and hence $\pi_k^* = \pi_{k+1}$ if $\tau_k^{(2)} > \xi_k$; otherwise, $f_{k+1}(\tau_{k+1}^{(2)}, \tau_k^{(2)}) \le 0$ and hence $\pi_k^* = \bar{\pi}_k$ by Lemma A.3 (a).

(c) If $\rho < \left(\sum_{\ell=1}^{k+1} p_\ell + \sqrt{(p_k + p_{k+1})p_{k+1}}\right)^{-1}$, then (EC.14) holds in the opposite direction for both $i = k$ and $i = k+1$. Then by Lemma A.3 (c), for $i, j \in \{k, k+1\}$ and $i \ne j$, there exists $\xi_j = h_i(\tau_i^{(2)}) > \tau_i^{(2)}$ such that $\pi_k^* = \pi_i$ if and only if $\tau_j^{(2)} > \xi_j$. Then,

$$\pi_k^* = \begin{cases} \pi_k, & \text{if } \tau_{k+1}^{(2)} > \xi_{k+1}; \\ \pi_{k+1}, & \text{if } \tau_k^{(2)} > \xi_k; \\ \bar{\pi}_k, & \text{otherwise.} \end{cases}$$

Furthermore, when $p_k \le p_{k+1}$, $h_k(\tau_k^{(2)})$ strictly increases in $\tau_k^{(2)}$ from Lemma A.3 (c). Hence, we can let $\tilde{\xi}_k \equiv h_k^{-1}(\tau_{k+1}^{(2)})$, where $h_k^{-1}(\cdot)$ is the inverse function of $h_k(\cdot)$, and $h_k^{-1}(\cdot)$ is also a strictly increasing function. Then, $\tau_{k+1}^{(2)} > \xi_{k+1} = h_k(\tau_k^{(2)})$ is equivalent to $h_k^{-1}(\tau_{k+1}^{(2)}) = \tilde{\xi}_k > \tau_k^{(2)}$. Hence,

$$\pi_k^* = \begin{cases} \pi_k, & \text{if } \tau_k^{(2)} < \tilde{\xi}_k; \\ \bar{\pi}_k, & \text{if } \tilde{\xi}_k \le \tau_k^{(2)} \le \xi_k; \\ \pi_{k+1}, & \text{if } \tau_k^{(2)} > \xi_k. \end{cases}$$

Finally, since $h_k(\tau_{k+1}^{(2)}) > \tau_{k+1}^{(2)}$, we have $\tau_{k+1}^{(2)} > h_k^{-1}(\tau_{k+1}^{(2)}) = \tilde{\xi}_k$. □

Proof of Corollary 2: We here only prove part (a); proofs of parts (b) and (c) are similar. If $C'_{k+1}(t)\mu_{k+1} \ge \max\{\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}}\}\mu_k$ for all $t \ge 0$, then for any non-negative random variable X, we have $E\left[C'_{k+1}(X)\right]\mu_{k+1} \ge \max\{\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}}\}\mu_k$ when the expectation exists. Furthermore, by Remark A.1, we have $\delta_i^{\gamma} = E[C'_i(U_i^{\gamma})]$. Hence,

$$\delta_{k+1}^{\pi_{k+1}}\mu_{k+1} = E\left[C'_{k+1}(U_{k+1}^{\pi_{k+1}})\right]\mu_{k+1} \ge \max\{\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}}\}\mu_k \ge \delta_k^{\pi_{k+1}}\mu_k,$$

which implies $C_{\pi_{k+1}} \le C_{\bar{\pi}_k}$ by Proposition 2 (b). Similarly, Proposition 2 (a) yields $C_{\bar{\pi}_k} \le C_{\pi_k}$. □

A.4. Supplemental material for Section 7

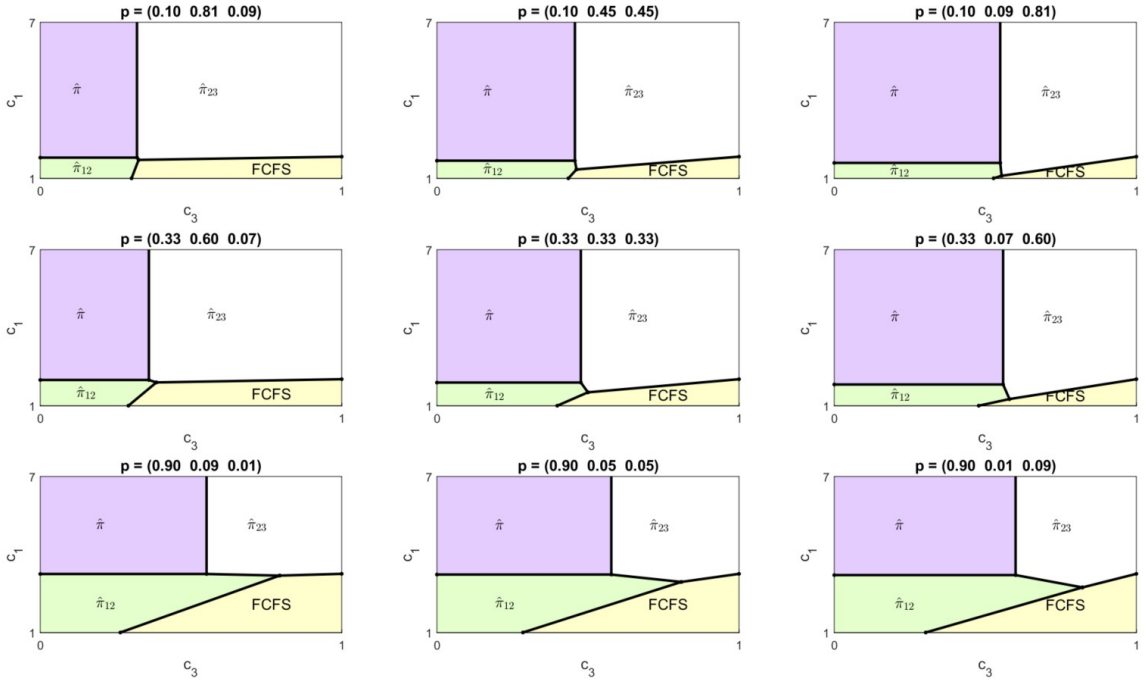

Figure A.1 **Optimal policy in Π^F with $K=3$, $\rho=0.7$, $C_i(t)=c_it^2$ for $i\in\{1,2,3\}$, $c_2=1$, $p=(p_1,p_2,p_3)$ indicated in each plot, and exponentially distributed service times with mean one.**

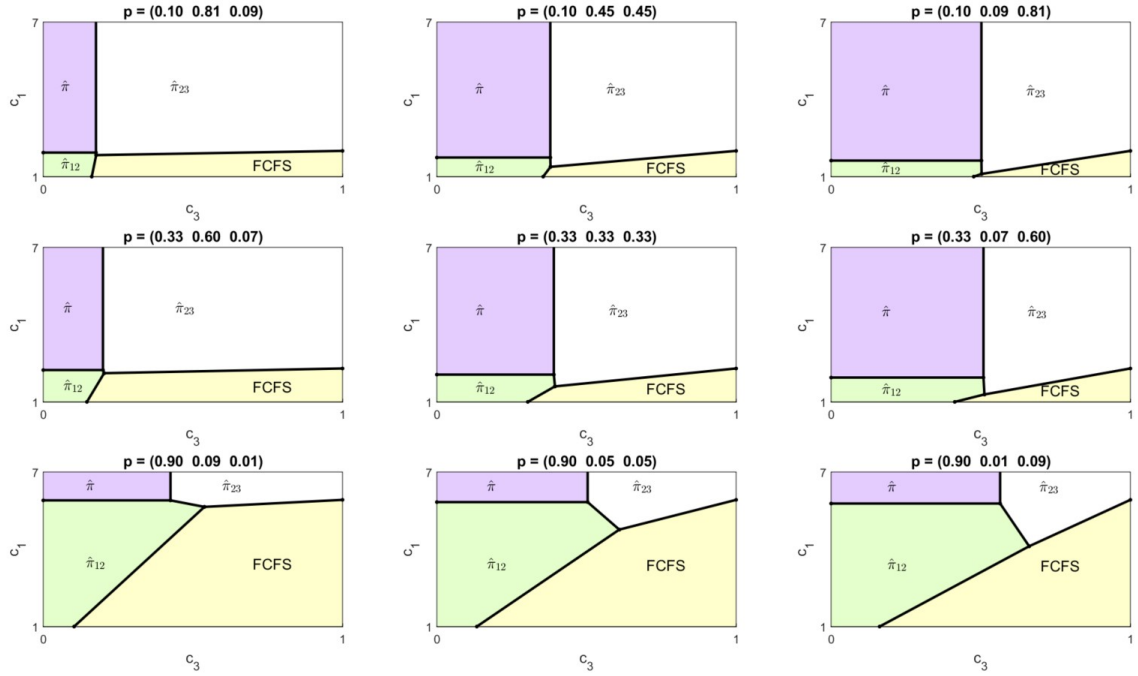

Figure A.2 **Optimal policy in Π^F with $K=3$, $\rho=0.9$, $C_i(t)=c_it^2$ for $i\in\{1,2,3\}$, $c_2=1$, $p=(p_1,p_2,p_3)$ indicated in each plot, and exponentially distributed service times with mean one.**

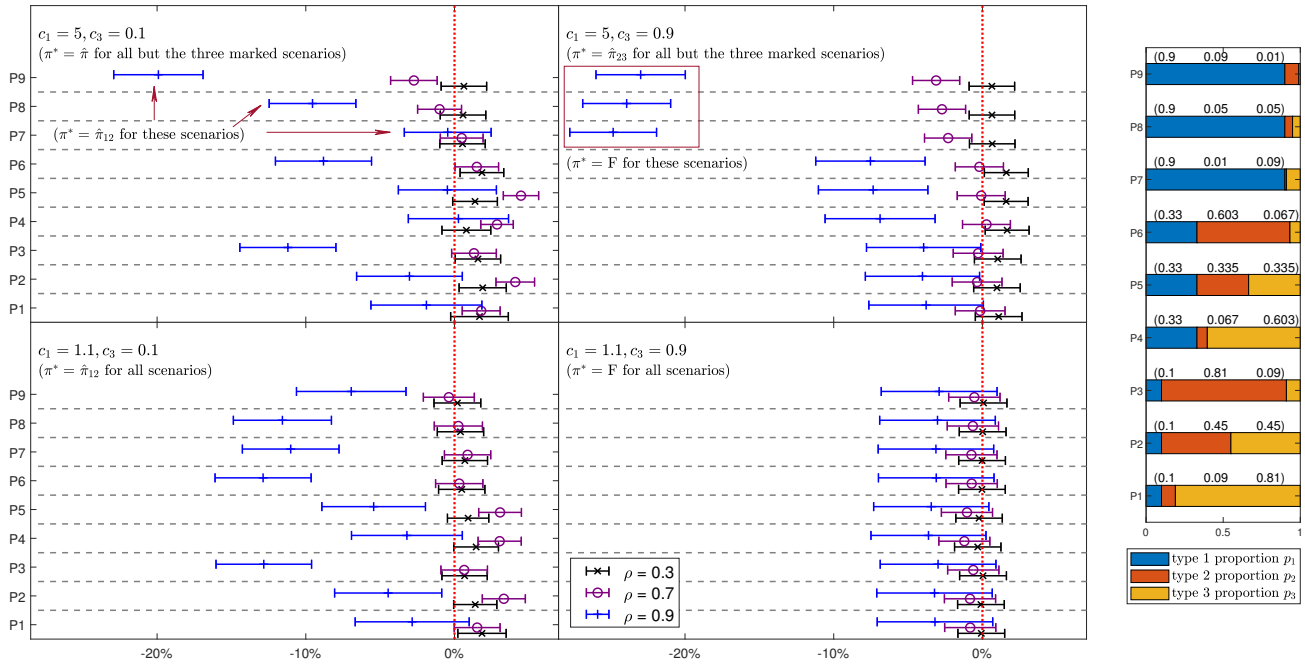

Figure A.3 **95% C.I. of the relative cost difference between G-$c\mu$ rule and π^* for the case with $K = 3$, where negative values indicate that G-$c\mu$ rule has a smaller cost than π^*, and the right-most graph shows the proportions of each type (p_1, p_2, p_3) for different scenarios in the simulation study**

Table A.4 **Threshold values R^{π_1} and R^{π_2} to characterize the best policy in Π^F for $K = 2$ that accompany numerical results of Section 7.2.**

		$\mu_1 = 5$		$\mu_1 = 1$		$\mu_1 = 0.2$	
ρ	p_1	R^{π_2}	R^{π_1}	R^{π_2}	R^{π_1}	R^{π_2}	R^{π_1}
0.3	0.1	0.375	1.260	0.532	1.612	0.805	2.558
	0.5	0.380	1.265	0.580	1.725	0.791	2.618
	0.9	0.390	1.240	0.620	1.88	0.794	2.670
0.7	0.1	0.240	1.590	0.307	1.831	0.508	2.552
	0.5	0.290	1.650	0.450	2.223	0.606	3.470
	0.9	0.390	1.970	0.546	3.255	0.628	4.202
0.9	0.1	0.105	1.850	0.169	2.000	0.347	2.560
	0.5	0.200	1.975	0.375	2.664	0.506	5.000
	0.9	0.390	2.880	0.500	5.918	0.540	9.312