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

Huiyin Ouyang, Nilay Tanık Argon, Serhan Ziya

Faculty of Business and Economics, The University of Hong Kong, Hong Kong; Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Received: January 11, 2018
Revised: August 14, 2019; August 27, 2020
Accepted: November 17, 2020
Published Online in Articles in Advance: April 29, 2021

https://doi.org/10.1287/mnsc.2020.3952

Copyright: © 2021 INFORMS

Abstract. 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 heterogeneous, so 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: Accepted by Baris Ata, stochastic models and simulation.
Funding: This work was supported by the National Science Foundation (NSF), Division of Civil, Mechanical and Manufacturing Innovation [Grants 1234212 and 1635574].

Keywords: priority assignment • priority queues • nonlinear waiting costs

1. Introduction

Many service systems prioritize their customers based on customer 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, that is, assigned a criticality level, and prioritized based on their triage category and arrival order. Another example is call centers, at which 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 60 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 because of 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 considers the possibility of nonlinear waiting costs imposes 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.

Although prioritization is prevalent in practice, in many cases, the policies in place are not based on careful statistical estimation of the waiting cost functions...
and are 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 patient survival, which can be seen as waiting cost, is not well understood (see Sacco et al. (2005), Jenkins et al. 2008, and the discussion on survival probability functions in Sun et al. 2017). Similarly, in other settings, such as 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, such as classifying customers into a few groups and prioritizing one group over the other without taking into account the state of the 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 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 that 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.) Although 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 these goals, we analyze an M/G/1 queueing system with two or more types of customers and each type characterized by a service time distribution and a waiting cost function, in which 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 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 in which certain classes of customers have priority over others are called priority queues.
The study of priority queues dates back to Cobham (1954), who consider a single-server Markovian queueing system (M/M/1) in which customers belong to multiple priority classes and the service is non-preemptive. For such a system, Cobham (1954) derives 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 advance the analysis of priority queues further, for example, 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 consider probabilistic priority policies, in which priorities are assigned randomly among different customer classes; for example, Katayama and Takahashi (1992) and Jiang et al. (2002) provide approximations for the delay performance under such policies.

When the waiting time of customers is penalized linearly with time, Cox and Smith (1961) establish the optimality of the well-known cμ-rule, which minimizes the long-run average waiting cost in an M/G/1 queue with multiple priority classes. According to the cμ-rule, customers with larger cμi index are assigned higher priority; cμi is the waiting cost per unit time, and μi is the service rate for type i customers. Following this seminal paper, optimality of the cμ-type policies is studied under various settings by Kakalik and Little (1971), Klimov (1974, 1979), Harrison (1975), Pinedo (1983), Nain (1989), Argon and Ziya (2009), and 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 cμ-type policies to research on achievable regions for optimal control of queueing systems, for example, Shanthikumar and Yao (1992) and Bertsimas (1995).

Although 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) show that, when waiting cost functions are increasing and convex, the optimal policy always serves customers that, when waiting cost functions are increasing and convex, the optimal policy always serves customers that belong to type i ∈ {1, 2, ..., K} with probability p_i > 0, where ∑_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 μ_i > 0 and nth moment τ^{(n)}_i > 0 for n ≥ 2. We define μ ≡ 1/ ∑_i=1^K (p_i/μ_i), τ^{(n)} ≡ ∑_i=1^K p_i τ^{(n)}_i, p_i ∝ λ_i/μ_i, and ρ ∝ λ/μ, which we call the traffic intensity, and we assume that ρ < 1 for stability. Each type i customer incurs a waiting cost C_i(t) when it is waiting in the queue. We assume that C_i(t) is first-order differentiable and nondecreasing in t for fixed i. (See Tables A.1, A.2, and A.3 in the Appendix for tables of notation used throughout this paper.)

For such a queueing system, we consider a set of policies Π that only includes nonidling and non-preemptive queueing policies that assign a fixed (deterministic) priority order to each type of customer. More specifically, Π consists of policies under which customers are ranked according to at most K priority orders, and any policy π ∈ Π satisfies the following properties: Let K^π be the number of distinct priority orders under policy π. Without loss of generality, let {1, 2, ..., K^π} denote the set of priority orders under policy π and assume that a smaller priority order represents a higher priority for service.

3. Model Description

Consider a single-server queueing system with K types of customers, in which K is an integer and 2 ≤ K < ∞. Customers arrive to the system according to a Poisson process with rate λ > 0, and each arriving customer belongs to type i ∈ {1, 2, ..., K} with probability p_i > 0, where ∑_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 μ_i > 0 and nth moment τ^{(n)}_i > 0 for n ≥ 2. We define μ ≡ 1/ ∑_i=1^K (p_i/μ_i), τ^{(n)} ≡ ∑_i=1^K p_i τ^{(n)}_i, p_i ∝ λ_i/μ_i, and ρ ∝ λ/μ, which we call the traffic intensity, and we assume that ρ < 1 for stability. Each type i customer incurs a waiting cost C_i(t) when it is waiting in the queue. We assume that C_i(t) is first-order differentiable and nondecreasing in t for fixed i. (See Tables A.1, A.2, and A.3 in the Appendix for tables of notation used throughout this paper.)

For such a queueing system, we consider a set of policies Π that only includes nonidling and non-preemptive queueing policies that assign a fixed (deterministic) priority order to each type of customer. More specifically, Π consists of policies under which customers are ranked according to at most K priority orders, and any policy π ∈ Π satisfies the following properties: Let K^π be the number of distinct priority orders under policy π. Without loss of generality, let {1, 2, ..., K^π} 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, \(1 \leq \kappa_i \leq K\).) 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 \(\Pi\) are also nonidling and nonpreemptive 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 that 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

\[
C_\pi = \lim_{t \to \infty} \frac{\sum_{s=1}^{K} \sum_{k=1}^{n_s(t)} C_i \left(V_{i,k}^{s,n_s(t)}\right)}{t}
\]  

(whenever the limit exists), where \(n_s(t)\) is the number of type \(i\) arrivals by time \(t\) and \(V_{i,k}^{s,n_s(t)}\) is the waiting time of the \(k\)th arriving type \(i\) customer under policy \(\pi\) and initial state \(s_0\). Our objective is to compare policies in \(\Pi\) in terms of their long-run average waiting costs in which smaller the long-run average cost, the better the policy. Let \(W_{i,k}^\pi\) denote the steady-state waiting time of a type \(i\) customer under policy \(\pi\). We show in the appendix that the limit in (1) exists and satisfies

\[
C_\pi = \lambda \sum_{i=1}^{K} p_i E\left[C_i(W_{i,k}^\pi)\right]
\]

if Assumption 1 holds.

**Assumption 1.** For \(\pi \in \Pi\) and all \(i = 1, 2, \ldots, K\), \(E[C_i(W_{i,k}^\pi)] < \infty\).

In this paper, we focus on policies in the subset \(\Pi^f \subset \Pi\), in which customers with the same priority order are served according to FCFS. These policies are of special interest because of their common use in practice. Furthermore, if \(C_i(\cdot)\) is a convex function (in the nonstrict sense) for all \(i = 1, 2, \ldots, K\), then it is sufficient to compare only policies in \(\Pi^f\) to find an optimal policy within \(\Pi\). More precisely, proposition 1 in Haji and Newell (1971) implies that, if \(C_i(\cdot)\) 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 \(C_{\pi^f} \leq C_\pi\). (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 for polynomial cost functions.

**Proposition 1.** Suppose that, for type \(i \in \{1, 2, \ldots, K\}\) customers, \(C_i(t) = \sum_{j=1}^{\ell_i} c_i^{(j)} t^j\), where \(j < \infty\) is the degree of the polynomial function \(C_i(t)\) and \(c_i^{(j)}\) are some real numbers such that \(C_i(t) \geq 0\) for all \(t \geq 0\). If \(p < 1\) and the first \(j + 1\) moments of service times for all customers are finite, then \(E[(W_i^\pi)^j] < \infty\) for \(\ell = 1, 2, \ldots, \ell_i\), and \(\gamma \in \Pi^f\), and hence, \(E[C_i(W_i^\pi)] < \infty\).

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 \(\Pi^f\) 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 \(\Pi^f\) 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 increases 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 \(\Pi^f\) corresponds to the ordered Bell number in number theory, which can be approximated by \(K!/(2(\log 2)^{K+1})\); see, for example, Gross (1962). This shows that it becomes much more difficult to compare all policies in \(\Pi^f\) as \(K\) increases.

Although we are not able to compare all policies in \(\Pi^f\) 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 predetermined.

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 \(\pi\) for \(i = 1, 2, \ldots, K\). For fixed \(k \in \{1, 2, \ldots, K - 1\}\), we compare \(\pi\) with two of its variants. We first define policy \(\pi_{k+1}\) to be a priority policy in which the priority orders of all types are the same as those under policy \(\pi\) except that the priority orders of type \(k\) and type \(k + 1\) customers are switched. We also define
policy $\pi_k$ to be a priority policy in which the priority orders of all types of customers are the same as those under policy $\pi$ except that type $k$ and type $k+1$ customers are pooled and served according to FCFS. Thus, policies $\pi_{k+1}$ and $\pi_k$ that are derived from policy $\pi$ treat all types the same way except for types $k$ and $k+1$: $\pi_{k+1}$ prioritizes type $k+1$ customers over type $k$, and $\pi_k$ does not differentiate between types $k$ and $k+1$ customers in terms of priority. For notational convenience, we also define policy $\bar{\pi}_k$, which is identical to policy $\pi$. Our next result provides necessary and sufficient conditions for the comparison of $\pi_k$, $\pi_{k+1}$, and $\pi_k$ under general cost structures.

**Proposition 2.** Suppose that Assumption 1 holds under policies $\pi_k$, $\pi_{k+1}$, and $\pi_k$ for fixed $k \in \{1, 2, \ldots, K-1\}$ and let

$$\delta_i^{(\gamma_1, \gamma_2)} = \frac{E[C_i(W_i^{(\gamma_1)})] - E[C_i(W_i^{(\gamma_2)})]}{E[W_i^{(\gamma_1)}] - E[W_i^{(\gamma_2)}]},$$

for $i \in \{k, k+1\}$, $\gamma_1, \gamma_2 \in \{\pi_k, \pi_{k+1}, \bar{\pi}_k\}$, and $\gamma_1 \neq \gamma_2$. Then, we have

a. $C_{\pi_k} \leq C_{\bar{\pi}_k}$ if and only if $\delta_i^{(\bar{\pi}_k, \pi_{k+1})} \leq \delta_i^{(\pi_k, \pi_{k+1})}$

b. $C_{\pi_k} \leq C_{\bar{\pi}_k}$ if and only if $\delta_i^{(\bar{\pi}_k, \pi_{k+1})} \geq \delta_i^{(\pi_k, \pi_{k+1})}$

c. $C_{\bar{\pi}_k} \leq C_{\pi_k}$ if and only if $\delta_i^{(\bar{\pi}_k, \pi_{k+1})} \geq \delta_i^{(\pi_k, \pi_{k+1})}$

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 $\pi_k$, $\pi_{k+1}$, 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 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 $\gamma_1$ to $\gamma_2$ 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 $\pi_{k+1}$ 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 $\pi_{k+1}$ to $\pi_k$ 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 $\Pi^r$ itself for $K = 2$. Indeed, as we show in Corollary A.1, we only need to compute $\delta_i^{(\gamma_1, \gamma_2)}$ for $i \in \{k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$ to find the best policy among $\pi_k$, $\pi_{k+1}$, and $\bar{\pi}_k$ using Proposition 2. Hence, in the rest of this paper, we simplify the notation by letting $\delta_i^{(\gamma_1, \gamma_2)}$ for $i \in \{k, k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$.

Furthermore, under certain assumptions on the structure of the waiting time functions, it is possible to obtain closed-form expressions for $\delta_i^{(\gamma_1, \gamma_2)}$ which we are using to derive the quadratic functions in Section 5. Finally, note that the computation of $\delta_i^{(\gamma_1, \gamma_2)}$ 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, for example, Ata and Tongarlik (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 regarding when to assign priorities when waiting costs are nonlinear.

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

$$M_i^{\gamma} = \frac{E[(W_i^{(\gamma)})^2] - E(W_i^{(\gamma)})^2}{E[W_i^{(\gamma)}] - E[W_i^{(\gamma)}]},$$

for $i \in \{k, k+1\}$ and $\gamma \in \{\pi_k, \pi_{k+1}\}$. (We drop the subscript from $W_i^{(\gamma)}$ for $i \in \{k+1\}$ because $W_i^{(\gamma)}$ and $W_i^{(\gamma)}$ are identical.) In words, $M_i^{\gamma}$ 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 $\gamma$ to $\pi_k$. As can be seen in Proposition 3, we need $M_i^{\gamma}$ to characterize the best policy among $\pi_k$, $\pi_{k+1}$, and $\bar{\pi}_k$ under quadratic costs. Closed-form expressions of $M_i^{\gamma}$ are given in Equations (A.8) and (A.9) and show that $M_i^{\gamma}$ 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, \( \delta_{i}^{(1)} \leq \delta_{j}^{(1)} \) for \( i, j \in \{k, k + 1\} \) and \( j \neq i \), where \( \delta_i^{(1)} = c_i^{(2)} M_i^{(1)} + c_i^{(1)} \) 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 \( \pi_k, \pi_{k+1} \), and \( \pi_k \) is characterized as follows:

a. If \( \delta_{k+1}^{(1)} \mu_k > \delta_{k+1}^{(1)} \mu_{k+1} \), then \( \pi_k \) is the best (and \( \pi_{k+1} \) is the worst).

b. If \( \delta_{k+1}^{(1)} \mu_{k+1} > \delta_{k+1}^{(1)} \mu_k \), then \( \pi_{k+1} \) is the best (and \( \pi_k \) is the worst).

c. Otherwise, that is, if \( \delta_k^{(1)} \mu_k \leq \delta_{k+1}^{(1)} \mu_k \), and \( \delta_{k+1}^{(1)} \mu_{k+1} \leq \delta_k^{(1)} \mu_k \), \( \pi_k \) is the best.

Proposition 3 is a generalization of the classical \( c_k \)-rule to the quadratic cost setting. (One can recover the \( c_k \)-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 \( c_k \)-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 \( \delta_i^{(1)} \) and \( \delta_j^{(1)} \) and part (c), we know that there is a nonempty region in which \( \pi_k \) 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 although FCFS is suboptimal under linear costs. The reason is that, when costs are linear, unlike in the nonlinear 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 nonlinear 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 nonpriority type, whereas this would not be the case in FCFS because of the randomization of order of arrivals of different types. Proposition 3 also states that serving both types \( k \) and \( k + 1 \) according to FCFS, that is, not giving priority between types \( k \) and \( k + 1 \), is never the worst policy. This can be best explained by the strict order between \( \delta_i^{(1)} \) and \( \delta_j^{(1)} \) for \( i, j \in \{k, k + 1\} \) and \( j \neq i \) because it implies that the expected marginal decrease in cost for type \( i \) by switching from \( \pi_{\bar{i}} \) to \( \pi_i \) is (strictly) less than the expected marginal increase in cost for the same type by switching from \( \pi_k \) to \( \pi_{\bar{k}} \). 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 in which \( c_k^{(1)} \mu_k = c_{k+1}^{(1)} \mu_{k+1} \), for example, when \( c_k^{(1)} = c_{k+1}^{(1)} = 0 \). (Argon et al. (2009) and Ata and Tongarlak (2013) are examples of work that studies 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 \( \pi_k, \pi_{k+1}, \) and \( \pi_k \) is characterized as follows: \( \pi_{k+1} \) is the best (and \( \pi_k \) is the worst) if \( \bar{c}_k^{(2)} \mu_k / (c_{k+1}^{(2)} \mu_{k+1}) < R^{\pi_{k+1}} / \mu_k \); \( \pi_k \) is the best (and \( \pi_{k+1} \) is the worst) if \( \bar{c}_k^{(2)} \mu_k / (c_{k+1}^{(2)} \mu_{k+1}) > R^{\pi_k} ; \) and \( \pi_k \) is the best if \( R^{\pi_{k+1}} / \mu_k \leq c_k^{(2)} \mu_k / (c_{k+1}^{(2)} \mu_{k+1}) \leq R^{\pi_k} / \mu_k \), where \( R' \equiv M_k^{(2)} / M_{k+1}^{(2)} \) for \( \gamma \in \{\pi_k, \pi_{k+1}\} \) and \( R^{\text{FCFS}} = R' \). 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 \( \pi_k, \pi_{k+1}, \) and \( \pi_k \) for quadratic cost functions with \( c_k^{(1)} \mu_k = c_{k+1}^{(1)} \mu_{k+1} \). In particular, it states that \( \pi_k \) is the best if \( \bar{c}_k^{(2)} \mu_k / (c_{k+1}^{(2)} \mu_{k+1}) \leq R^{\pi_{k+1}} / \mu_k \) and \( \bar{c}_k^{(2)} \mu_k / (c_{k+1}^{(2)} \mu_{k+1}) > R^{\pi_k} / \mu_k \). Using the work conservation law (see, e.g., Kleinrock 1965), \( R' \) 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 \( \gamma \) to \( \pi_k \) 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 \( \bar{c}_k^{(2)} \mu_k / (c_{k+1}^{(2)} \mu_{k+1}) \) given in Corollary 1, over which one policy is better than the others, change with system parameters such as 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 \( c_i^{(2)} \) 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 \( \rho \), proportion of type 1 customers \( p_1 \), and the service time distributions under quadratic cost functions \( C_i(t) = c_i^{(2)} t^2 \), where \( c_i^{(2)} > 0 \) and \( i \in \{1, 2\} \). First, note that, when there are two types of customers in the system, \( \pi_i \) 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

\[ \text{Corollary 1.} \]
provides a complete characterization of the best policy in \( \Pi \): \( \pi_1 \) is the best if \( c_1(2) \mu_1 / (c_2(2) \mu_2) > R_{\pi_1} \), \( \pi_2 \) 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 \( \rho \) (or, equivalently, with \( \lambda \)) and \( p_1 \).

Figures 1 and 2 show how the optimal policy shifts from \( \pi_2 \) to FCFS first and then to \( \pi_1 \) 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 in which FCFS is the best policy enlarges as \( \lambda \) increases and that both \( R_{\pi_1} \) and \( R_{\pi_2} \) monotonically change with \( \lambda \), but they are not necessarily increasing or decreasing in all cases. (By the symmetry between the two types and the rightmost plot in Figure 1, we know that there is also a case in which both \( R_{\pi_1} \) and \( R_{\pi_2} \) increase with \( \rho \).) We also notice from Figure 2 that \( R_{\pi_1} \) and \( R_{\pi_2} \) 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 \( R_{\pi_1} \) and \( R_{\pi_2} \) and also provide their limits in heavy traffic.

**Proposition 4.**

a. \( R_{\pi_1} \) increases in \( \lambda \) if \( \tau_2(2) \mu_2 \geq \tau_1(2) \mu_1(1 - 2p_1)/(2 - 2p_1) \).

b. \( R_{\pi_2} \) decreases in \( \lambda \) if \( \tau_1(2) \mu_1 \geq \tau_2(2) \mu_2(1 - 2p_2)/(2 - 2p_2) \).

c. As \( \lambda \to \mu = (p_1/\mu_1 + p_2/\mu_2)^{-1} \), we have

\[
R_{\pi_1} \to \frac{p_1 \mu_1^{-1} + 2p_2 \mu_2^{-1}}{p_2 \mu_2^{-1}} \text{ and } R_{\pi_2} \to \frac{p_1 \mu_1^{-1}}{2p_1 \mu_1^{-1} + p_2 \mu_2^{-1}}.
\]

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, \( R_{\pi_1} \) and \( R_{\pi_2} \) both increase in \( p_1 \) (and, hence, decrease in \( p_2 \)).

---

**Figure 1.** (Color online) Regions of Optimality for FCFS, \( \pi_1 \), and \( \pi_2 \) as a Function of \( \rho \) (or, Equivalently, \( \lambda \)) Under Exponential Service Times and Quadratic Waiting Costs with \( C_i(t) = c_i(t)^2 \) for \( t \geq 0 \) and \( i \in \{1, 2\} \)

**Figure 2.** (Color online) Regions of Optimality for FCFS, \( \pi_1 \), and \( \pi_2 \) as a Function of \( p_1 \) Under Exponential Service Times and Quadratic Waiting Costs with \( C_i(t) = c_i(t)^2 \) for \( t \geq 0 \) and \( i \in \{1, 2\} \)
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; 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 $\tau^{(2)}_1 \mu_i$ values for the two customer types are relatively close, specifically, one is not more than twice as large as the other, then the region in which FCFS is the best gets larger as the traffic intensity grows, and the optimality regions for the priority policies get smaller: from parts (a) and (b) of Proposition 4, if $1/2 < \tau^{(2)}_1 \mu_1 / (\tau^{(2)}_2 \mu_2) < 2$, then $R^{n_1}$ increases and $R^{n_2}$ decreases with $\lambda$.

- The region in which 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, $R^{n_1}$ increases ($R^{n_2}$ decreases) with $\lambda$. This also can be 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 Tongarlik (2013), and Ata and Peng (2018), in which 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, and the optimality region for prioritizing the other type enlarges: this directly follows from Proposition 5 and can be also seen from the leftmost 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 $c^{(2)}_i/\mu_i$ 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, $R^{n_1} \to 2$ and $R^{n_2} \to 0$ as $\rho_1 \to 0$, whereas $R^{n_1} \to \infty$ and $R^{n_2} \to 1/2$ as $\rho_1 \to 1$. 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, that is, type $3-i$, should be served first if $c^{(2)}_{3-i}/\mu_{3-i} > 2c^{(2)}_i/\mu_i$, and otherwise, FCFS should be applied.

In this section, we also compare the values of $R^{n_1}$ and $R^{n_2}$ under two different service time distributions with the same means to observe the effects of service time distributions (or, equivalently, second moments in the quadratic cost case). For $\gamma \in \{n_1, n_2\}$, let $R^{\gamma}_{exp}$ and $R^{\gamma}_{det}$ denote the values of $R^{\gamma}$ under exponential and deterministic service times for all customers, respectively.

**Proposition 6.**

a. $R^{n_1}_{exp} \geq R^{n_1}_{det}$ if and only if $\mu_1 \geq \mu_2(1-\rho_1)/(2-\rho_1)$.

b. $R^{n_2}_{exp} \leq R^{n_2}_{det}$ if and only if $\mu_1 \leq \mu_2(2-\rho_2)/(1-\rho_2)$.

Proposition 6 implies that, if $\mu_2/\mu_1 \in (1-\rho_2)/(2-\rho_2) \sim (2-\rho_1)/(1-\rho_1)$, then $R^{n_2}_{exp} \leq R^{n_1}_{det}$ at $R^{n_1}_{exp}$ 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 $c^{(2)}_1/c^{(2)}_2$ 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 also have higher variance regardless of whether FCFS or a strict fixed priority policy is in place. Nevertheless, because of the convexity of the waiting cost functions, the impact is 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 $R^{n_1}_{exp} \leq R^{n_1}_{det}$ and $R^{n_2}_{exp} \leq R^{n_2}_{det}$, which implies that, under deterministic service times, $n_2$ (prioritizing the faster type) is preferred for a wider range of values of $c^{(2)}_1/c^{(2)}_2$, and $n_1$ (prioritizing the slower type) is preferred for a narrower range of values of $c^{(2)}_1/c^{(2)}_2$ than that under exponential service times.

We 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 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 which 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 nonidling 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 backward under a policy compared with FCFS and 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 considers 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 $\Pi$ as can be verified by the work conservation law. Hence, minimizing the variance of the steady-state waiting times within $\Pi$ is equivalent to minimizing its second moment, which corresponds to letting $C_\pi(t) = t^2$ ($t \geq 0$) for all $\pi$ in our formulation. Then, we can use Corollary 1 to prove the following result.

Proposition 7. Suppose that $\lambda, \mu_i, \tau^{(2)}_i$, and $\tau^{(3)}_i$ are finite and $\mu = \mu > \lambda$ 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 $\pi^*_k$ 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 $\rho \geq (\sum_{i=1}^{k-1} p_i + \sqrt{p_k + p_{k+1}p_k})^{-1}$, then $\pi^*_k = \bar{\pi}_k$.

b. If $\rho < (\sum_{i=1}^{k-1} p_i + \sqrt{p_k + p_{k+1}p_k})^{-1}$, then there exists a threshold $\xi_k > \tau^{(2)}_{k+1}$ such that

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

c. If $\rho < (\sum_{i=1}^{k+1} p_i + \sqrt{p_k + p_{k+1}p_k})^{-1}$, then there exist thresholds $\xi_k > \tau^{(2)}_{k+1}$ and $\tilde{\xi}_k < \tau^{(2)}_{k+1}$ such that

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

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 $\rho$ 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., $1 \leq p_{k+1}/p_k < (p_k + p_{k+1})/2 - 1$); 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 $\rho \geq (1 + \sqrt{\min(p_1, p_2, \ldots, p_K)})^{-1}$, that is, 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) and 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, because service rates are the same for types $k$ and $k + 1$, comparison of $\bar{\tau}^{(2)}_k$ and $\bar{\tau}^{(2)}_{k+1}$ 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; 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 $\bar{\xi}_k$ 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 provide 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 $\pi_k$, $\pi_{k+1}$, and $\pi_k$, 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 $\pi_k$, $\pi_{k+1}$, and $\pi_k$.

a. If $C_{k+1}(t) \geq \max(\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}})\mu_k / \mu_{k+1} \text{ for all } t \geq 0,$

then $C_{k+1} \leq C_k \leq C_{k+1}.$

b. If $C_{k+1}(t) \leq \min(\delta_k^{\pi_k}, \delta_k^{\pi_{k+1}})\mu_k / \mu_{k+1} \text{ for all } t \geq 0,$

then $C_{k+1} \leq C_k \leq C_{k+1}.$

c. If $\delta_k^{\pi_k} \mu_k / \mu_{k+1} \leq C_{k+1}(t) \leq \delta_k^{\pi_{k+1}} \mu_k / \mu_{k+1} \text{ for all } t \geq 0,$

then $C_{k+1} \leq C_k \leq C_{k+1},$ and $C_{k+1} \leq C_{k+1}.$

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 $\pi_k$, $\pi_{k+1}$, and $\pi_k$. 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.

6.1. Quadratic Cost for One Type

Suppose type $k$ customers are known to have a quadratic cost function, that is, $C_k(t) = c_k^2 + c_k^1 t + c_k^0$ for $c_k^0 \geq 0$ and $c_k^2 > 0$, but we do not know the exact form of $C_k(t)$. In this case, we have $\delta_k^{\pi_k} = c_k^2 M_k^\gamma + c_k^1$, where $M_k^\gamma$ 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, that is, $C_{k+1}(t)$ is at most $\delta_k^{\pi_{k+1}} \mu_k / \mu_{k+1}$, 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, that is, $C_{k+1}(t)$ 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 $C_k(t)$ 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 $\lambda$ (see (A.8)–(A.10)), 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 $\lambda$ becomes larger. Indeed, both $\delta_k^{\pi_k}$ and $\delta_k^{\pi_{k+1}}$ go to infinity as $\lambda$ approaches $\mu$ for $K = 1$, 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 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.

6.2. Linear Cost for at Least One Type

Suppose that type $k$ customers are known to have a linear cost function, that is, $C_k(t) = c_k t$ for $t \geq 0$ and $c_k > 0$, but we do not know the exact form of $C_k(t)$. Then, we have $\delta_k^{\pi_k} = \delta_k^{\pi_{k+1}} = c_k / \mu_k$, and by Corollary 2, if $C_{k+1}(t) \geq c_k \mu_t / \mu_{k+1}$ for all $t \geq 0$, then $C_{k+1} \leq C_k \leq C_{k+1}$, and if $C_{k+1}(t) \leq c_k \mu_t / \mu_{k+1}$ for all $t \geq 0$, then $C_{k+1} \leq C_k \leq C_{k+1}$. 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) $c_k \mu_t / \mu_{k+1}$, 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 $C_t(t) = c_j t$ for $t \geq 0$, $c_j > 0$, and $j = 2, 3, \ldots, K$. If $C_j(t) = \max_{i=2,..,K} c_i / \mu_t$ for all $t \geq 0$, then Corollary 2 implies that type $j$ should receive the highest priority, and all other types should follow the $c_j$-rule. Similarly, if $C_j(t) \leq \min_{i=2,..,K} c_i / \mu_t$ for all $t \geq 0$, then type $j$ should receive the lowest priority while all other types follow the $c_j$-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, then our result can provide a simple sufficiency condition for the optimality of prioritizing this type over all other types with linear costs. As an example, suppose that $C_j(t) = h_j^3 - h_j^2 + h_j t$ for $t \geq 0$, where $h_1, h_2, h_3$ are positive constants such that $C_1(t) > 0$, that is, $h_2 < 3h_3$. 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 conduct a numerical study on a system with three types as presented in Section 7.1. We also conduct 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 \( K = 3 \), in which 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 \( C_i(t) = c_i t^2 \) for \( i = 1, 2, 3 \). (For notational convenience, we drop the superscript \( c \) 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 13 policies in \( \Pi^E \) when \( K = 3 \) as discussed at the beginning of Section 4. Instead of computing and comparing the long-run average costs for all 13 policies to identify the best policy \( \pi^* \) in \( \Pi^E \), 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 \( c \); see the discussion in the paragraph following Corollary 1. Hence, out of the six policies in \( \Pi^E \) 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 \( \Pi^E \) that pools types \( k \) and \( \ell \) for \( 1 \leq k < \ell \leq 3 \), the waiting cost function for the pooled group of customers is \( \hat{C}_{k\ell}(t) = c_{k\ell} t^2, t \geq 0 \), where \( c_{k\ell} = (p_k c_k + p_l c_l)/(p_k + p_l) \).

Then, because \( c_{12} > c_3 \) 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 \( \Pi^E \), we numerically compute and compare the long-run average costs for the remaining six policies to find \( \pi^* \) 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 \( \pi^* \) 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 prioritize 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 Figures A.1 and A.2 in the Appendix.)

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

i. For \( k \in \{1, 2\} \) and \( \ell = k + 1 \), \( C_{\hat{n}} \leq C_{\hat{h}_k} \) if and only if

\[
\frac{c_k}{c_{k+1}} \geq \frac{\bar{\rho}_{k-1}(\bar{\rho}_{k+1} - 1) + \bar{\rho}_k - 1}{\bar{\rho}_k(\bar{\rho}_{k+1} + \bar{\rho}_k - 1) - 1};
\]

ii. \( C_{\text{FCFS}} \leq C_{\hat{h}_{12}} \) if and only if

\[
\frac{c_{12}}{c_3} \leq \frac{2 - \rho_1 - \rho_2 + (1 - \rho)(1 - \rho_1 - \rho_2)}{2 - \rho_1 - \rho_2 - \rho};
\]

iii. \( C_{\text{FCFS}} \leq C_{\hat{h}_{23}} \) if and only if

\[
\frac{c_{1}}{c_{23}} \leq \frac{2 - \rho_1 + (1 - \rho)(1 - \rho_1)}{2 - \rho_1 - \rho}.
\]

Figure 3. (Color online) Optimal Policy in \( \Pi^E \) with \( K = 3 \), \( p_i = 1/3 \), and \( C_i(t) = c_i t^2 \) 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
Figure 3 provides plots of optimality regions determined partially by bounds $\Gamma_1, \Gamma_2, \Gamma_{12}$, and $\Gamma_{23}$ provided. More specifically, the upper leftmost corner is where $\hat{\pi}$ is better than $\tilde{\pi}_{12}$ and $\tilde{\pi}_{23}$ (determined by lines $c_1 = \Gamma_1$ and $c_3 = \Gamma_2^{-1}$, respectively), and the lower rightmost corner is where FCFS is better than $\tilde{\pi}_{12}$ and $\tilde{\pi}_{23}$ (determined by lines $c_1 = \Gamma_{12}(1 + p_2/p_1)c_3 - p_2/p_1$ and $c_1 = \Gamma_{23}(p_3 + p_3c_1)/(p_2 + p_3)$, respectively). By using numerical comparisons, we also find that $\hat{\pi}$ and FCFS are indeed the best in $\Pi_F$ in these respective regions. The optimality regions for $\tilde{\pi}_{12}$ and $\tilde{\pi}_{23}$ are found numerically in Figure 3. (Our analytical results do not provide comparisons between FCFS and $\hat{\pi}$ or between $\tilde{\pi}_{12}$ and $\tilde{\pi}_{23}$. ) We make the following observations from Figures 3, A.1, and A.2, which extend most of our analytical observations for the case with two types of customers (i.e., the optimality regions for policies $\pi^*$ are prioritized 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 Figure A.3 in the Appendix). Assume that service times for type $i \in \{1, 2\}$ customers are exponentially distributed with rate $\mu_i$, where $\mu_2$ is fixed at one per unit time, and the cost functions take the form $C_{1}(t) = c_1t^2$ and $C_{2}(t) = t^2$, $t \geq 0$. 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 $\Pi^0$ by computing the values of $R_{\pi^*}$ and $R_{\pi^0}$ in Corollary 1 (reported in Table A.4 in the Appendix): $\pi_2$ has the smallest cost if $c_1 \mu_1 < R_{\pi^0}$, $\pi_1$ has the smallest cost if $c_1 \mu_1 > R_{\pi^*}$, and FCFS has the smallest cost if $R_{\pi^0} \leq c_1 \mu_1 \leq R_{\pi^*}$. To obtain the long-run average cost under the G-$\mu$ rule (denoted by $C_{*}$), we built a simulation model (using Simio 11 simulation software), in which we computed a priority index for each customer in the queue and assigned nonpreemptive 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, determined by the G-$\mu$ rule, is given by $2c_i \mu_i t$. We ran 100 independent replications of length 60,000 minutes for each scenario and truncated the first 6,000 minutes based on a warmup period analysis. We report the mean relative cost difference by using the G-$\mu$ rule over the best fixed priority policy, that is, $\left(C_{*} - C_{\pi^*}\right) \times 100/C_{\pi^*}$ (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 $\pi^*$ and the G-$\mu$ rule are different; the comparison is in favor of $\pi^*$ for a positive C.I. and the G-$\mu$ 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-$\mu$ rule is insignificant in most scenarios, especially when the traffic intensity is light or moderate, and the G-$\mu$ rule may have smaller costs in scenarios under heavy traffic. In several scenarios, for example, $c_1 = 5, \mu_1 = 5, p_1 = 0.9$ with traffic intensity 0.7 or 0.9 (when $\pi^* = \pi_1$) and $c_1 = 0.1, \mu_1 = 0.2, p_1 = 0.1$ with traffic intensity 0.7 (when $\pi^* = \pi_2$), 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, in which 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 is 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 $c_1 = 5, \mu_1 = 0.2, p_1 = 0.9, \rho = 0.9$ 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, the 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, for example, 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 study 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, that is, 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 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 compare these fixed priority policies (including FCFS) in terms of their long-run average performance and derive several managerial insights by focusing mostly on the case with quadratic waiting costs.

It is well known that, if all customers have linear waiting costs, then only the product of the rates of service and waiting cost 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 conclude 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 find that the heterogeneity of the population is determined by the population mix (i.e., proportions of each customer type in the population) and the first three moments of service times as well as cost parameters. Furthermore, we show that the arrival rate has a direct effect on the decision to prioritize or not. Specifically, we observe that the parameter region in which FCFS is best enlarges as the arrival rate increases. Hence, haphazardly replacing the 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 find that one should be especially cautious with prioritizing types that constitute a large proportion of the population mix. One setting in which it would be safe to replace the 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 by-product 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 shows that FCFS is the fairest policy when the population is homogenous in terms of service time variability. For a population with heterogeneous service-time variability, we show 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 implies that FCFS is better than any policy that groups customer types into two priority classes if the traffic intensity is larger than $1/(1 + \sqrt{p})$, where $p$ is the smallest proportion of any type in the population. However, if the traffic intensity is not heavy, then we show that prioritizing the type with smaller service time variance could actually be fairer than FCFS.

Because the focus of this work is on the use (or misuse) of simple but popular fixed priority policies in systems with nonlinear waiting costs, we mostly exclude 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 $G$-cµ 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 provide an exploratory analysis to facilitate interest on this research question by conducting a simulation study that compares the best fixed priority policy with the $G$-cµ rule under quadratic waiting costs. We find that the $G$-cµ 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 of the waiting cost function, performs similarly or even slightly better than the $G$-cµ rule. We believe that more research is needed to support these claims, especially given that the $G$-cµ rule is not always an optimal dynamic policy. Studying randomized priority policies, in which 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).

**Acknowledgments**

The authors thank the department editor, the associate editor, and the referees whose comments and suggestions greatly improved this paper.

**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.

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

**Lemma A.1.** For policies $\pi_k$, $\pi_{k+1}$, and $\pi_3$ that are defined in Section 4, we have
Table A.1. Notation for System Parameters

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda )</td>
<td>Arrival rate of all customers</td>
</tr>
<tr>
<td>( p_i )</td>
<td>Probability that an arriving customer belongs to type ( i )</td>
</tr>
<tr>
<td>( \mu_i )</td>
<td>Service rate of type ( i ) customers</td>
</tr>
<tr>
<td>( r_i^{(n)} )</td>
<td>( n )th moment of service time of type ( i ) customers for ( n \geq 2 )</td>
</tr>
<tr>
<td>( \rho_i )</td>
<td>Traffic intensity for type ( i ) customers, that is, ( \lambda p_i/\mu_i )</td>
</tr>
<tr>
<td>( \rho )</td>
<td>Traffic intensity for all customers, that is, ( \sum_{i=1}^{K} \rho_i )</td>
</tr>
<tr>
<td>( \mu )</td>
<td>Service rate for a random customer, that is, ( (\sum_{i=1}^{K} p_i/\mu) ) ( -1 )</td>
</tr>
<tr>
<td>( \gamma^{(n)} )</td>
<td>( n )th moment of service time for a random customer, that is, ( \sum_{i=1}^{K} p_i r_i^{(n)} ) for ( n \geq 2 )</td>
</tr>
<tr>
<td>( C_i(\cdot) )</td>
<td>Waiting cost function for type ( i ) customers</td>
</tr>
<tr>
<td>( C(\cdot) )</td>
<td>Long-run average waiting cost under policy ( \gamma )</td>
</tr>
<tr>
<td>( K^\gamma )</td>
<td>Number of priority orders under policy ( \gamma )</td>
</tr>
<tr>
<td>( \gamma(\cdot) )</td>
<td>Probability that an arriving customer has priority order ( j ) under policy ( \gamma ), i.e., ( \sum_{i=1}^{K} \gamma(i) = p_i )</td>
</tr>
<tr>
<td>( \rho_j^{(\gamma)} )</td>
<td>Service rate of customers with priority order ( j ) under policy ( \gamma )</td>
</tr>
<tr>
<td>( \rho_j )</td>
<td>Traffic intensity for customers with priority order ( j ) under policy ( \gamma )</td>
</tr>
<tr>
<td>( \gamma_j )</td>
<td>Probability that the length of a busy period in an M/G/1 queue is lower than ( j ) times of customers with priority order ( j ) under policy ( \gamma ) for ( j = 1, \ldots, K^\gamma )</td>
</tr>
<tr>
<td>( \Lambda^\gamma_k )</td>
<td>Arrival rate of customers with priority order ( k ) or smaller under policy ( \gamma ), that is, ( \lambda \sum_{i=1}^{k} \rho_i^{(\gamma)} )</td>
</tr>
</tbody>
</table>

Proof of Lemma A.1.

a. The distribution function for the steady-state waiting times of customers with priority order \( p \) under policy \( \gamma \) is given by equation 17 in Takács (1964) as follows:

\[
P\left[ W^\gamma_p \leq x \right] = \int_0^x \sum_{j=0}^{\infty} e^{-\Lambda^\gamma_{j+1}} \frac{(\Lambda^\gamma_{j+1})^j}{j!} B^\gamma_{pj}(x-y) \ dW^\gamma_{sp}(y),
\]

(A.1)

where \( B^\gamma_{pj}(x) \) 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 \( \Lambda^\gamma_{p(j)} \) and service time distribution \( S^p_{j(p)}(x) \) in which there are \( j \geq 0 \) customers initially in the system; \( B^\gamma_{pj}(x) = 1 \) for all \( x \geq 0 \) and \( j \geq 0 \); and \( W^\gamma_{sp}(y) \) 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 policy \( \gamma \) are pooled and served according to FCFS regardless of their actual priority order under policy \( \gamma \), and their service times are i.i.d. with distribution \( S^p_{j(p)}(x) \), and all other customers are served according to \( \gamma \).

For \( p = 1, 2, \ldots, k-1 \), it is easy to see that \( \Lambda^\gamma_{sp} = \Lambda^\gamma_{sp+1} = \Lambda^\gamma_{sp+2} \), \( S^p_{2p}(x) = S^p_{2p+1}(x) = S^p_{2p+2}(x) \), and hence, \( B^\gamma_{sp}(x) = B^\gamma_{sp+1}(x) = B^\gamma_{sp+2}(x) \) for all \( x \geq 0 \) and \( j \geq 0 \), and \( W^\gamma_{sp}(y) = W^\gamma_{sp+1}(y) = W^\gamma_{sp+2}(y) \) for all \( y \geq 0 \). Similarly, for \( p = k + 2, \ldots, K \), we have \( \Lambda^\gamma_{sp} = \Lambda^\gamma_{sp+1} = \Lambda^\gamma_{sp+2} \), \( S^p_{2p-1}(x) = S^p_{2p}(x) = S^p_{2p+2}(x) \), and

\[ B^\gamma_{sp}(x) = B^\gamma_{sp+1}(x) = B^\gamma_{sp+2}(x) \] for all \( x \geq 0 \) and \( j \geq 0 \), and \( W^\gamma_{sp}(y) = W^\gamma_{sp+1}(y) = W^\gamma_{sp+2}(y) \) for all \( y \geq 0 \).
Indicated in Each Plot and Exponentially Distributed Service Times with Mean One. The result follows directly from equation 68 of Takács (1964).

**Definition A.1** (Di Crescenzo 1999). Let \( X \) and \( Y \) be two nonnegative random variables with \( X \leq_{st} Y \) and \( E[X] < E[Y] < \infty \). Then, \( Z = \Psi(X, Y) \) is a random variable with probability density function

\[
f_Z(z) = \frac{f_X(x) - f_Y(x)}{E[Y] - E[X]}, \quad x \geq 0,
\]

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 nonnegative 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, \( E[g(Z)] = F_G(z) \) is finite and

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

Lemma A.2 presents a probabilistic analogue of the mean value theorem, in which \( Z \) 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 sense) between \( X \) and \( Y \). For example, when \( X \) and \( Y \) are exponential random variables with distinct rates, \( Z = X + Y \) (see example 3.1 in Di Crescenzo 1999).

**A.2. Proofs of Results and Supplemental Material for Sections 3 and 4**

**A.2.1. Proof of Equivalence of Equations (1) and (2).** The long-run average cost in (1) can be written as

\[
C_n = \lim_{t \to \infty} E[X_i(t)] = \lim_{i \to \infty} n_i(t) - t
\]

\[
= \sum_{i=1}^{K} \lambda_i \lim_{t \to \infty} \sum_{k=1}^{n_i(t)} C_i \left( \frac{V_i^{(x_0)}}{n_i(t)} \right)
\]

\[
E[Y_n(t)] = E[Y_n(t)], \tag{A.3}
\]

which follows from the fact that \( n_i(t) \) is a Poisson process with rate \( \lambda_i \). In the following, we prove that, for \( i \in \{1, 2, \ldots, K\} \) when \( E[C_i(W_i^n)] \) is finite,

\[
\lim_{n \to \infty} \sum_{k=1}^{n} C_i \left( \frac{V_i^{(x_0)}}{n_i(t)} \right) = E[C_i(W_i^n)], \tag{A.4}
\]

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

In the remainder of this proof, we drop the superscripts \( \pi \) and \( x_0 \) for notational convenience and let \( T_{ik}, S_{ik}, \) and \( D_{ik} \) be


Table A.2. Notation for Random Variables (RV), Their Cumulative Density Functions (CDF), and LSTs

<table>
<thead>
<tr>
<th>RV</th>
<th>CDF</th>
<th>LST</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{i0}^{(k)} )</td>
<td>—</td>
<td>—</td>
<td>Waiting time of the ( k )th arriving type ( i ) customer under policy ( \gamma ) and initial state ( x_0 )</td>
</tr>
<tr>
<td>( W_i^\gamma )</td>
<td>—</td>
<td>—</td>
<td>Steady-state waiting time of a type ( i ) customer under policy ( \gamma )</td>
</tr>
<tr>
<td>( W_i^{(j)} )</td>
<td>—</td>
<td>—</td>
<td>Steady-state waiting time for a customer with priority order ( j ) under policy ( \gamma )</td>
</tr>
<tr>
<td>( S_i(\cdot) )</td>
<td>—</td>
<td>—</td>
<td>Service time for type ( i ) customers</td>
</tr>
<tr>
<td>( S_i^{(j)}(\cdot) )</td>
<td>—</td>
<td>—</td>
<td>Service time for customers with priority order ( j ) under policy ( \gamma ), that is, ( S_i^{(j)}(x) = \sum_{t=0}^{\infty} p_i^{(j)} S_i(x) / p_i^{(j)} )</td>
</tr>
<tr>
<td>( S_{\lambda k} )</td>
<td>—</td>
<td>—</td>
<td>Service time for customers with priority order ( k ) or smaller under policy ( \gamma ), that is, ( S_{\lambda k}(x) = \sum_{t=0}^{\infty} p_i^{(\lambda)} S_i(x) / \sum_{t=0}^{\infty} p_i^{(\lambda)} )</td>
</tr>
<tr>
<td>( S(\cdot) )</td>
<td>—</td>
<td>—</td>
<td>Service time for a random customer, that is, ( S(x) = \sum_{i=1}^{\infty} p_i S_i(x) )</td>
</tr>
<tr>
<td>( B_i(\cdot) )</td>
<td>—</td>
<td>—</td>
<td>Length of a busy period in an M/G/1 queue with arrival rate ( \lambda_{\gamma p=1} ) and service time distribution ( S_{\gamma p=1}(\cdot) ) for ( p = 2, 3, \ldots, K' ) under policy ( \gamma ) in which there are ( j \geq 1 ) customers initially in the system</td>
</tr>
<tr>
<td>( U_i^{\gamma,12} )</td>
<td>—</td>
<td>—</td>
<td>( \Psi(W_{i1}^{(\gamma)}, W_{i2}^{(\gamma)}) ) for ( i \in {k, k+1} ), ( \gamma_1, \gamma_2 \in {\pi_1, \pi_2, \pi_3} ), and ( \gamma_1 \neq \gamma_2 ), where ( \Psi(\cdot) ) is defined in Definition A.1</td>
</tr>
<tr>
<td>( T_{\bar{R}_i} )</td>
<td>—</td>
<td>—</td>
<td>( U_i^{(\lambda)} ) for ( \gamma \in {\pi_1, \pi_2, \pi_3} ) and ( i \in {k, k+1} )</td>
</tr>
<tr>
<td>( T_{\bar{R}_i} )</td>
<td>—</td>
<td>—</td>
<td>Length of a busy period in which only type ( j \in {1, 2, \ldots, K} ) ( {l} ) customers arrive</td>
</tr>
<tr>
<td>( \bar{B}(\cdot) )</td>
<td>—</td>
<td>—</td>
<td>Length of a busy period starting from an empty and idle system for an M/G/1 queue under any policy in ( \Pi )</td>
</tr>
</tbody>
</table>

the arrival, service, and departure times of the \( k \)th type \( i \) customer, respectively, under policy \( \pi \) and initial state \( x_0 \). Then, \( V_{i0} = D_i - T_{i0} - S_{i0} \) is the queue waiting time for this customer. Note that \( \{V_{i0}, k = 1, 2, \ldots\} \) for each \( i \in \{1, 2, \ldots, K\} \) is a delayed regenerative process with \( n \)th regeneration happening at \( N_{i0} \) for \( n = 0, 1, 2, \ldots, \) where \( N_{i0} = 1 \) and \( N_{i0} = \min\{k : k > N_{i0-1}, V_{i0} = 0\} \). Note also that, for each \( i \), \( \{C_i(V_{i0}), k = 1, 2, \ldots\} \) is a regenerative process with the same regeneration epochs as \( \{V_{i0}\} \). Then, by theorem 13 of chapter 2 and last paragraph of page 93 in Wolff (1989), (A.4) holds for \( i \in \{1, 2, \ldots, K\} \) if \( \sum_{k=1}^{\infty} |C_i(V_{i0})| < \infty \) with probability one, \( E[N_{i2} - N_{i1}] < \infty \), and \( E[|\sum_{k=1}^{\infty} C_i(V_{i0})|] < \infty \). We next complete the proof by showing that these three conditions hold.

When \( \rho < 1 \), the system returns 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).

Figure A.2. (Color online) Optimal Policy in \( \Pi_1^f \) with \( K = 3 \), \( \rho = 0.9 \), and \( c_i(t) = c_i t^2 \) 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
This implies that $N_{i1} < \infty$ with probability one, $N_{i2} - N_{i1} < \infty$ with probability one, $V_{i1} < \infty$ for any $i$ and $k$ with probability one, and $E[N_{i2} - N_{i1}] < \infty$. At last, by theorem B.(5) in El-Taha and Stidham (1999), $E[\sum_{i_1=1}^{n_i} |C_i(V_{i0})|] = E[C_i(W_{i0})]E[N_{i2} - N_{i1}]$ is finite under the assumption that $E[C_i(W_{i0})]$ is finite. □

A.2.2. 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 $W_j^* \leq W_{j0}^*$ for any policy $\gamma \in \Pi^F$ by a similar interchange argument as the proof of Lemma A.1(b). Then, for $1 \leq \ell < \infty$, by theorem 1A.3(a) of Shaked and Shanthikumar (2007), we have $E(W_{j0}^*) \leq E(W_{j0}^{\gamma^*})$ for any policy $\gamma \in \Pi^F$. Note that Assumptions 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 $E(W_{j0}^{\gamma^*})$ is finite for all $\ell = 1, 2, \ldots, j$. Note also that
\[
\frac{d}{ds} W_{j0}^{\gamma^*} (s) = \lambda (1 - p_j) (1 - B_{k}(s)) + s),
\]
where $D_{k}(s) = (1 - p_j) / s - \lambda (1 - S(s))$ and $B_{k}(s)$ is the unique solution to $B_{k}(s) = S_{k}(s) + \lambda (1 - p_j) (1 - B_{k}(s))$ for $s > 0$ and $\lim_{s \rightarrow 0} B_{k}(s) = 0$, and $S_{k}(s) = (S(s) - p_j S(s))/ (1 - p_j)$. Then, using Faa di Bruno’s formula (see, e.g., Theorem 2 of Roman 1980), (A.5) is finite if $d(W_{j0}^{\gamma^*}) / dr \bigg|_{r=0}$ and $ddB_{k}(s) / ds \bigg|_{s=0}$ are finite for all $n \leq \ell$, that is, if the nth moment of $W_{j0}$ and $T_{j0}$ are finite.

When $\rho < 1$, we can obtain the nth moment of $W_{j0}$ as (see, e.g., Gross et al. 2008)
\[
E[(W_{j0})^\eta] = \frac{\lambda}{1 - \rho} \sum_{i=1}^{n_i} \left( \frac{p_j}{\ell + 1} \right)^{i} E[(W_{j0})^{\ell + 1}],
\]
where $E[(W_{j0})^{\ell + 1}]$ is the $(\ell + 1)$st moment of visit time of a randomly picked customer. Hence, $E[(W_{j0})^{\eta}]$ is finite if $\rho < 1$ and the first $n + 1$ moments of service times of all customers are finite. Besides, from Theorem 1 of Graharamani 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, $E[(W_{j0}^{\gamma^*})^\ell]$ is finite if $\rho < 1$ and the first $\ell + 1$ moments of service times are finite. □

A.2.3. 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
\[
N_{i} - N_{i1} = \lambda \sum_{j=1}^{n_i} p_j \left( E[C_i(W_{j0}^*)] - E[C_i(W_{j0}^\gamma)] \right) + \lambda p_{k+1} \left( E[C_i(W_{j0}^*)] - E[C_i(W_{j0}^\gamma)] \right) + \lambda p_{k+1} \left( E[C_{k+1}(W_{j0}^*)] - E[C_{k+1}(W_{j0}^\gamma)] \right),
\]
where the last equation follows from Lemma A.1(a). Then, $N_{i} - N_{i1} \leq \mu_{k+1}$ if and only if
\[
\mu_{k+1} E[C_{k+1}(W_{j0}^*)] - E[C_{k+1}(W_{j0}^\gamma)] \leq \mu_{k} E[C_{k}(W_{j0}^*)] - E[C_{k}(W_{j0}^\gamma)].
\]

According to the work conservation law (see, e.g., Kleinrock 1965), we have
\[
\sum_{i=1}^{K} \mu_i E[W_{j0}^{\gamma^*}] / \mu_i = \sum_{i=1}^{K} \mu_i E[W_{j0}^{\gamma}] / \mu_i,
\]
and by Lemma A.1(a), we have $E[W_{j0}^{\gamma}] = E[W_{j0}^{\gamma^*}]$ for $j \in \{1, 2, \ldots, K\} \setminus \{k, k+1\}$. Hence,
\[
\mu_{k+1} E[W_{j0}^{\gamma^*}] - E[W_{j0}^{\gamma}] / \mu_{k+1} = \mu_{k} E[W_{j0}^{\gamma}] - E[W_{j0}^{\gamma^*}] / \mu_{k},
\]
which is positive by Lemma A.1(c). Dividing (A.6) by (A.7) completes the proof. □

Remark 1. For $i \in \{k, k+1\}$, $\gamma_i, \gamma_j \in \{n_i, n_{i1}, n_{i2}\}$, and $\gamma_1 \neq \gamma_2$, let $U_{i1}^{(\gamma_1)} \equiv \Psi(W_{j0}^{\gamma_1}, W_{j0}^{\gamma_2})$, where $\Psi(\cdot, \cdot)$ is defined in Definition A.1. Note that $U_{i1}^{(\gamma_1)}$ is well defined for $i \in \{k, k+1\}$ and $\gamma_1, \gamma_2 \in \{n_i, n_{i2}, n_{i1}\}$ because $E[W_{j0}^{\gamma_1}] < E[W_{j0}^{\gamma_2}] < E[W_{j0}^{\gamma_1}]$ from Lemma A.1(c) and because of the stochastic ordering provided in Lemma A.1(b). Then, from Lemma A.2, we have $\delta_{i}^{(\gamma_1)} = E[C(\ell_{i}^{(\gamma_1)}))]$ for $i \in \{k, k+1\}$ and $\gamma_i \in \{n_i, n_{i1}, n_{i2}\}$.

In an immediate corollary to Proposition 2, we provide necessary and sufficient conditions for the optimality of $n_i$, $n_{i1}$, and $n_{i2}$ within the set of these three policies, for which we only need to calculate the values of $\delta_{i}^{(\gamma_1)} \equiv \delta_{i}^{(\gamma_1)}$ for $i \in \{k, k+1\}$ and $\gamma \in \{n_k, n_{k1}\}$.

Corollary 1. Let $I(n_{i1}) \equiv 0$ and $I(n_k) \equiv (\delta_{i1}^{(\gamma_1)} - \delta_{i1}^{(\gamma_2)}) / (\rho_{k-1} - \rho_{k})$ for $i \in \{k, k+1\}$ and $i \neq j$. Then, the policy with the largest (smallest) value of $I(\gamma)$ for $\gamma \in \{n_k, n_{k1}, n_{k2}\}$ has the lowest (highest) long-run average cost among these three policies.

Proof of Corollary 1. For $i \in \{k, k+1\}$, $I(n_{i1}) \geq I(n_{i2}) \Leftrightarrow \delta_{i1}^{(\gamma_1)} \leq \delta_{i1}^{(\gamma_2)} \Leftrightarrow C_{n_{i1}} \leq C_{n_{i2}}$, which follows from Proposition 2(a) and (b). Furthermore, for $i \in \{k, k+1\}$ and $i \neq j$, we get
\[
\delta_{i}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) + \delta_{i1}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) = \frac{\delta_{i}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) + \delta_{i1}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right)}{2 \rho_{k-1} - \rho_{k} + \rho_{k+1}},
\]
which follows from Lemma A.1(c). Then, by Proposition 2(c),
\[
I(n_{i2}) \geq I(n_{i1}) \Leftrightarrow \mu_{k} \left[ \delta_{i}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) + \delta_{i1}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) \right] \geq \mu_{k} \left[ \delta_{i}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) + \delta_{i1}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) \right] \geq \mu_{k} \left[ \delta_{i}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) + \delta_{i1}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) \right] \geq \delta_{i}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) + \delta_{i1}^{(\gamma_1)} \left( \rho_{k-1} - \rho_{k} \right) \geq C_{n_{i1}} \leq C_{n_{i2}}.
\]
and $E(W_i^n)^2$ 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):

$$M_i^n = \frac{1}{\bar{p}_{k-1}} \left[ \frac{2\gamma^{(3)}}{3\gamma^{(2)}} \left( \frac{\lambda}{p_1} \sum_{l=1}^{i-1} p_r \tau_l^{(2)} + \tau_{i-1}^{(2)} \right) \right] + \left[ \frac{\lambda}{p_1} \sum_{l=1}^{i-1} p_r \tau_l^{(2)} + \tau_{i-1}^{(2)} \right] \hat{p}_{k-1} - \hat{p}_i - \hat{p}_j,
$$

(8.8)

$$M_j^n = \frac{1}{\bar{p}_{k-1}} \left[ \left( \frac{2\gamma^{(3)}}{3\gamma^{(2)}} \left( \frac{\lambda}{p_1} \sum_{l=1}^{i-1} p_r \tau_l^{(2)} + \tau_{i-1}^{(2)} \right) \right) \right] + \left[ \frac{\lambda}{p_1} \sum_{l=1}^{i-1} p_r \tau_l^{(2)} + \tau_{i-1}^{(2)} \right] \hat{p}_{k-1} - \hat{p}_i - \hat{p}_j
$$

(8.9)

We next show that $\delta_i^{(j)} < \delta_i^{(j)}$ when $H_i^{(j)} > 0$ for $i \in \{k, k+1\}$ and $j = 2k + 1 - i$ by showing that $M_i^n < M_j^n$. By subtracting the indices of $i$ and $j$ in (8.9) and subtracting (8.8), we have

$$M_i^n - M_j^n = \frac{2\gamma^{(3)}}{3\gamma^{(2)}} \left( \frac{\lambda}{p_1} \sum_{l=1}^{i-1} p_r \tau_l^{(2)} + \tau_{i-1}^{(2)} \right) \left( \frac{1}{\hat{p}_{k-1} - \hat{p}_i - \hat{p}_j} \right)
$$

(10.8)


\[\text{Hence, } R_{n+1}^i < R_{n}^i. \square\]

### A.3.3. Proof of Proposition 4

We first show that, for $i \in \{1, 2\}$, $\frac{\partial}{\partial \lambda} \left[ \frac{M_i^n}{M_3^n} \right] < 0$ if

$$\frac{\gamma i_{t-M_3}^{(3)}}{\lambda} \geq \frac{1 - \rho_i}{2(1 - \rho_i)}.
$$

(11.1)

We define $G_i(\lambda)$ for $i \in \{1, 2\}$ as $G_i(\lambda) = \frac{2\gamma^{(3)}}{3\gamma^{(2)}} \frac{\lambda}{p_1} + \frac{\lambda p_t}{1 - p_1}$. Then, (8.8) and (8.9) reduce to

$$M_i^n = G_i(\lambda) \left( \tau_i^{(2)} + \tau_{i-1}^{(2)} \right), \quad M_j^n = \frac{2 - \rho_i}{1 - \rho_i} G_i(\lambda) \left( \tau_i^{(2)} + \tau_{i-1}^{(2)} \right), \quad \text{for } i \in \{1, 2\}.
$$

Then, we have

$$G_i(\lambda) \left( \tau_i^{(2)} + \tau_{i-1}^{(2)} \right) \hat{p}_{i-M_3}^{(3)} = \frac{\partial}{\partial \lambda} \left( \frac{M^n_i}{M_{3-n}^i} \right) = - \left( \frac{2\gamma^{(3)}}{3\gamma^{(2)}} \right) \frac{\lambda}{p_1} \frac{G_i(\lambda)}{\left( \frac{2\gamma^{(3)}}{3\gamma^{(2)}} \right) G_i(\lambda) + \tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}} < 0
$$

if and only if

$$\tau_i^{(2)} \hat{p}_i \left( 1 - \rho_i \right) \geq \frac{\tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}}{\left( 1 - \rho_i \right)}
$$

(11.2)

Because, for $i \in \{1, 2\}$, $G_i(\lambda) = \frac{\tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}}{\left( 1 - \rho_i \right)} > 0$. Note that the right-hand side of (11.2) is greater than

$$\frac{\tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}}{\left( 1 - \rho_i \right)} > \frac{\tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}}{\left( 1 - \rho_i \right)} + \frac{\tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}}{\left( 1 - \rho_i \right)}
$$

(11.2)

Thus, a sufficient condition for (11.2) to hold is

$$\tau_i^{(2)} \hat{p}_i \left( 1 - \rho_i \right) \left( \frac{\tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)}}{\left( 1 - \rho_i \right)} \right) < \tau_i^{(2)} \hat{p}_i \left( 1 - \rho_i \right) + \tau_{i-1}^{(2)} \hat{p}_{i-M_3}^{(3)} \left( \frac{\rho_i}{1 - \rho_i} \right)
$$

which reduces to (11.1).

Now, note that $R_{n+1}^i = M_3^n / M_i^n$ increases in $\lambda$ if and only if $M_i^n / M_3^n$ decreases in $\lambda$. Then, by letting $i = 1$ in (11.1),
we obtain part (a). Similarly, by letting $i = 2$ in (A.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,

$$
\lim_{\lambda \to \mu} \frac{M_i^\alpha}{M_i^\beta} = \lim_{\lambda \to \mu} \left( \frac{2 - \rho_i}{\lambda - \mu} \right) G_i(\lambda) + \tau_i^{(2)} \mu_i = \lim_{\lambda \to \mu} \left( \frac{2 - \rho_i}{\lambda - \mu} \right) G_i(\lambda)
$$

$$
= \lim_{\lambda \to \mu} \frac{1 - \rho_i}{\lambda - \mu} \frac{p_{3,i}/\mu_i - p_{2,i}/\mu_i + 2p_{3,i}/\mu_i}{p_{3,i}/\mu_i - p_{2,i}/\mu_i}.
$$

Then, letting $i = 1$ and $i = 2$ provides the limits for $R^{\pi_1}$ and $R^{\pi_2}$ as $\lambda \to \mu$. □

**A.3.4. Proof of Proposition 5.** Because the first three moments of service times do not depend on type, we drop the subscript from $\mu_i$, $\tau_i^{(2)}$, and $\tau_i^{(3)}$. Then, we have

$$
R^{\pi_1} = \frac{(2 - \rho_1)M + \tau_1^{(2)}\mu_1}{M + \tau_1^{(2)}\mu_1} \frac{1 - \rho_1}{1 - \rho}, \quad R^{\pi_2} = \frac{(2 - \rho_2)M + \tau_2^{(2)}\mu_2}{M + \tau_2^{(2)}\mu_2} \frac{1 - \rho_2}{1 - \rho},
$$

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

$$
\frac{\partial R^{\pi_1}}{\partial \rho_1} = \frac{\partial M}{\partial \rho_1} \frac{1 - \rho_i}{1 - \rho} > 0 \quad \text{and} \quad \frac{\partial R^{\pi_2}}{\partial \rho_2} = \frac{-\partial M}{\partial \rho_1} \frac{1 - \rho_i}{1 - \rho} < 0. \quad \Box
$$

**A.3.5. 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, $\tau_i^{(3)} = 1/\mu_i$ and $\tau_i^{(2)} = 1/\mu_i^2$ for $i \in \{1, 2\}$, and hence, we have

$$
R^{\pi_1}_{\exp} = \frac{(2 - \rho_1)N_{\exp}^{(1)} + \frac{1}{\mu_1}}{N_{\exp}^{(1)} + \frac{1}{\mu_1}}, \quad R^{\pi_2}_{\exp} = \frac{(2 - \rho_2)N_{\exp}^{(2)} + \frac{1}{\mu_2}}{N_{\exp}^{(2)} + \frac{1}{\mu_2}}.
$$

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

$$
N_{\exp}^{(i)} = \frac{p_1/\mu_1^3 + p_2/\mu_2^3}{p_1/\mu_1^3 + p_2/\mu_2^3} + \frac{\lambda(p_1/\mu_1^2 + p_2/\mu_2^2)}{1 - \rho},
$$

$$
N_{\det}^{(i)} = \frac{2(p_1/\mu_1^3 + p_2/\mu_2^3)}{3(p_1/\mu_1^3 + p_2/\mu_2^3)} + \frac{\lambda(p_1/\mu_1^2 + p_2/\mu_2^2)}{1 - \rho},
$$

where $N_{\exp}^{(i)} > N_{\det}^{(i)}$. Taking the difference of $R^{\pi_1}_{\exp}$ and $R^{\pi_2}_{\det}$, we have

$$
R^{\pi_1}_{\exp} - R^{\pi_2}_{\det} = \left( \frac{1}{\mu_2} \frac{(2 - \rho_2)}{1 - \rho} - \frac{1}{\mu_1} \right) \left( N_{\exp}^{(1)} - N_{\det}^{(1)} \right) \left( N_{\det}^{(1)} + \frac{1}{\mu_1} \right) \left( N_{\exp}^{(1)} + \frac{1}{\mu_1} \right).
$$

Hence, $R^{\pi_1}_{\exp} \geq R^{\pi_2}_{\det}$ 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 $R^{\pi_2}_{\exp} - R^{\pi_1}_{\det}$.

**A.3.6. Proof of Proposition 7.** We first provide Lemma A.3 to facilitate the proof of Proposition 7.
Lemma A.3. Suppose that $\lambda, \mu_i, \tau_i^{(2)}$, and $\tau_i^{(3)}$ are finite and $\mu_i = \mu > \lambda$ for all $i = 1, 2, \ldots, K$. Then,

a. $\pi_i^* = \pi_i$ if and only if $f_i(\tau_i^{(2)}, \tau_i^{(3)}) > 0$ for $i \in \{k, k+1\}$ and $j = 2k + 1 - i$, where for $x_2 > 0$,

$$f_i(x_1, x_2) = \langle \tilde{p}_{k-i} \tilde{p}_{k+1} - (p_k + p_{k+1})(\tilde{p}_{k+1} - \rho_i) \rangle x_2 - \rho_i (p_k + p_{k+1}) + \frac{\tilde{p}_{k-1} \tilde{p}_{k+1}}{\rho_k - \rho_i} - \frac{1}{\rho_k - \rho_i} x_1, \quad \text{(A.13)}$$

\[ -\tilde{p}_{k-1} \tilde{p}_{k+1} + \frac{2\tau_i^{(3)}}{3 \mu \rho_k \rho_k(\rho_k + \rho_{k+1})^2} + \frac{1}{\rho_k - \rho_i} x_1. \]

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

$$\rho \geq \frac{1}{\Sigma_{i=1}^{k+1} p_i^2 + (\sqrt{p_k + p_{k+1}})^2}, \quad \text{(A.15)}$$

which is feasible for $\rho < 1$ when $p_k + p_{k+1} > \Sigma_{i=k+2}^{K} p_i^2$ and $p_{2k+1} > \Sigma_{i=k+2}^{K} \rho_i^2 (p_k + p_{k+1})$.

c. For $i \in \{k, k+1\}$, if (A.15) 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 $p_i \leq p_{2k+1}$, then $h_i(x_1)$ strictly increases in $x_1$.

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

$$M_i^{\ast} - M_i^{\circ} = \frac{1}{\rho_i} - \frac{1}{\rho_k - \rho_i} = \frac{\mu (\tau_i^{(2)} - \tau_i^{(3)})}{\tilde{p}_{k-1} \tilde{p}_{k+1}} x_1 \times \frac{2\tau_i^{(3)}}{3 \mu \rho_k \rho_k(\rho_k + \rho_{k+1})^2} + \frac{1}{\rho_k - \rho_i} x_1 > 0$$

if and only if $f_i(\tau_i^{(2)}, \tau_i^{(3)}) > 0$ for $i \in \{k, k+1\}$ and $j = 2k + 1 - i$. (b) By (A.13), $f_i(x_1, x_2) < 0$ for any $x_1, x_2 > 0$ if $\tilde{p}_{k-1} \tilde{p}_{k+1} - (p_k + p_{k+1})(\tilde{p}_{k+1} - \rho_i) \leq 0$, which reduces to (A.14). For $i \in \{k, k+1\}$, (A.14) could hold for $\rho < 1$ only if its right-hand side is less than one, that is,

$$\left( \frac{\tilde{p}_{k-1} \tilde{p}_{k+1}}{p_k + p_{k+1}} \right) > \frac{1}{\rho_k - \rho_i} x_1 \times \frac{2\tau_i^{(3)}}{3 \mu \rho_k \rho_k(\rho_k + \rho_{k+1})^2} + \frac{1}{\rho_k - \rho_i} x_1 > 0$$

which holds only if we have $p_k + p_{k+1} > \Sigma_{i=k+2}^{K} p_i^2$.

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

$$f_i(x_1, x_1) < \tilde{p}_{k-1} \tilde{p}_{k+1} - \frac{\tilde{p}_{k-1} \tilde{p}_{k+1}}{p_k + p_{k+1}} \rho_i x_1.$$
Lemma A.3(c), for
in the opposite direction for both
A.4. Supplemental Material for Section 7
\[ \frac{p_i}{k} + \frac{p_k}{p_i} \]
by parts (a) and (b) of Lemma A.3) and holds in the oppo-
site direction for \( i = k + 1 \). Now, let \( \xi_k \equiv h_k(t_k^{(2)}(\cdot)) \), where
\( h_k(\cdot) \) is defined in Lemma A.3(c) and \( \xi_k > t_k^{(2)}(\cdot) \). Then,
f_{k+1}(t_k^{(2)}, t_k^{(2)}) > 0, and hence, \( \pi_k^* = \pi_{k+1} \) if \( t_k^{(2)} \geq \xi_k \); otherwise,
f_{k+1}(t_k^{(2)}, t_k^{(2)}) \leq 0, and hence, \( \pi_k^* = \pi_k \) by Lemma A.3(a).

(c) If \( \pi > (\sum_{i=1}^{k+1} p_i + \sum_{i=1}^{k+1} p_k) t_k^{(2)}(\cdot)^{-1} \), then (A.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 \( i \neq j \), there exists \( \xi_j = h_j(t_j^{(2)}(\cdot)) > t_j^{(2)}(\cdot) \) such that \( \pi_k^* = \pi_j \) if and only if \( t_j^{(2)}(\cdot) > \xi_j \). Then,
\[ \pi_k^* = \begin{cases} \pi_k, & \text{if } t_k^{(2)} > \xi_k; \\ \pi_{k+1}, & \text{if } t_k^{(2)} \geq \xi_k; \\ \pi_k, & \text{otherwise}. \end{cases} \]

Furthermore, when \( p_k \leq p_{k+1}, h_k(t_k^{(2)}(\cdot)) \) strictly increases in \( t_k^{(2)} \) from Lemma A.3(c). Hence, we can let \( \xi_k \equiv h_k^{-1}(\tau_k^{(2)}(\cdot)) \), 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, \( \tau_k^{(2)} > \tau_k^{(1)} + \xi_k = h_k^{-1}(\tau_k^{(2)}) \) is equivalent to \( h_k^{-1}(\tau_k^{(2)}) = \xi_k > \tau_k^{(2)} \). Hence,
\[ \pi_k^* = \begin{cases} \pi_k, & \text{if } t_k^{(2)} < \xi_k; \\ \pi_{k+1}, & \text{if } \xi_k \leq t_k^{(2)} \leq \xi_k; \\ \pi_{k+1}, & \text{if } t_k^{(2)} > \xi_k. \end{cases} \]

Finally, because \( h_k(t_k^{(2)}) > h_k(t_k^{(1)}) \), we have \( t_k^{(2)} > h_k^{-1}(t_k^{(2)}) = \xi_k \).

\[ \text{A.4. Supplemental Material for Section 7} \]

Table A.4. Threshold Values \( R^{(2)} \) and \( R^{(3)} \) to Characterize the Best Policy in \( \Pi^f \) for \( K = 2 \)
That Accompany the Numerical Results of Section 7.2

<table>
<thead>
<tr>
<th>( \mu_1 = 5 )</th>
<th>( \mu_1 = 1 )</th>
<th>( \mu_1 = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>( R^{(2)} )</td>
<td>( R^{(3)} )</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.375</td>
</tr>
<tr>
<td>0.3</td>
<td>0.380</td>
<td>1.265</td>
</tr>
<tr>
<td>0.9</td>
<td>0.390</td>
<td>1.240</td>
</tr>
<tr>
<td>0.7</td>
<td>0.240</td>
<td>1.590</td>
</tr>
<tr>
<td>0.5</td>
<td>0.290</td>
<td>1.650</td>
</tr>
<tr>
<td>0.9</td>
<td>0.390</td>
<td>1.970</td>
</tr>
<tr>
<td>0.9</td>
<td>0.105</td>
<td>1.850</td>
</tr>
<tr>
<td>0.9</td>
<td>0.200</td>
<td>1.975</td>
</tr>
<tr>
<td>0.9</td>
<td>0.390</td>
<td>2.880</td>
</tr>
</tbody>
</table>

References


Argon NT, Ziya S (2009) Priority assignment under imperfect in-


