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

Huiyin Ouyang, Nilay Tanık Argon, Serhan Ziya

January 11, 2018

Abstract

For queueing systems with multiple customer types differing in service time distributions and costs for waiting, it is known that giving priority to one type over others minimizes the long-run average waiting costs when waiting is penalized linearly in time. However, when waiting costs are nonlinear, which is typically a more reasonable depiction of reality, it is not clear whether policies that ignore the type information such as the first-come-first-serve policy (FCFS) should be replaced with type-based priority policies. To shed some light on to this problem, we study a single-server queueing system with two types of customers under static queueing policies that use information on customers' types and order of arrival. Our main theorem ranks the type-based priority policies and FCFS according to their long-run average waiting costs under nonlinear cost functions. We then apply this result to polynomial cost functions and generate insights into when prioritization is advantageous. For example, we find that when customers are similar in terms of their service time distributions, then the parameter region where FCFS is more preferable over type-based priority policies under quadratic costs increases with traffic intensity. We also conduct a numerical study to compare the best static policy with a well-known dynamic policy that requires information on the current waiting times of customers. We find that the best static policy performs comparably with (sometimes even better than) this dynamic policy except when the traffic is heavy and it is not clear which type should receive priority.

1 Introduction

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

Despite significant progress, however, this literature has important gaps from both academic and practical point of view largely due to the assumptions imposed on the waiting costs for analytical tractability. Specifically, an overwhelming majority of the articles assume 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 our literature review in Section 2). On the other hand, the work that considered the possibility of nonlinear waiting costs imposed some other restrictions on the system such as the requirement that the system be operating under heavy traffic and the waiting cost function being convex. More importantly, the policies proposed (e.g., the generalized $c\mu$ rule, which is first proposed by Van Mieghem (1995)) are somewhat sophisticated requiring the system to keep track of the arrival time of each customer in the queue and having a complete knowledge of the waiting cost function, which may pose a challenge in practice.

While prioritization is prevalent in practice, in many cases, the policies in place are not based on careful statistical estimation of the waiting cost functions but mostly based on some rough analysis of limited data, and the service providers' past experience and beliefs about who needs the service more urgently or whose long wait would be more detrimental for the system. For example, prioritization of patients in emergency departments or in the aftermath of mass-casualty events is very common in practice yet the precise nature of the effect of passage of time on the patient survival, which can be seen as waiting cost, is not well understood (see Jenkins et al. (2008), Sacco et al. (2005) and the discussion on survival probability functions in Sun et al. (2017)). Similarly, in other settings like healthcare clinics and call centers, there is very limited work on the estimation of waiting costs. Nevertheless, this does not stop providers from implementing prioritization policies that they believe to be improving system performance. They also usually stick with simple policies like classifying customers into two groups and prioritizing one group over the other.

Given the fact that waiting cost functions are not known precisely, choosing a simple prioritization policy (if one needs to be chosen) is reasonable. But the question remains as to whether 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-serve (FCFS) policy, which is at least largely perceived to be fair? A provider who uses prioritization without knowing the precise form of the waiting cost functions is in fact implicitly assuming a certain relationship between the waiting cost functions for different classes. But what are these implicit assumptions? One of the two main goals of this article is to provide some answers to these questions, which we do by comparing the performance of FCFS with those of priority policies under cost functions that are not necessarily linear.

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

In the pursuit of the goals stated above, we analyze an M/G/1 queueing system with two customer classes, where each type is characterized by a service time distribution and a waiting cost function and the waiting cost function for at least one type is nonlinear. (Complete model description is provided in Section 3.) We formally define our objective to be to determine the best "static policy" that minimizes the long-run average cost. By a static policy, we mean a policy that determines the order customers are taken into service using information *only* on customer type identities and their rank with respect to the arrival times *but not* on the number of customers in the system and their current waiting times. These policies include priority policies that give priority to a certain type of customers and non-priority policies like FCFS.

We first provide our main result, which completely characterizes the best policy among the three commonly used static policies, namely, F, PF_1 and PF_2, where F is short for FCFS and

PF_i denotes the *type-based* priority policy that prioritizes type i customers and employs FCFS within each type for $i = 1, 2$, under general waiting cost functions (see Section 4). (For convex cost functions, based on earlier results from the literature, we can show that it is sufficient to consider only these three static policies to find the optimal static policy.) In Section 5, we use our main theorem to provide a clearer characterization of the best policy when waiting cost functions are polynomial, which leads to several managerial insights. For example, through this analysis, we find that, if waiting cost functions are quadratic, unlike in the linear-cost case, the optimal static policy depends on the traffic intensity and the proportion of each type in the customer population. We also obtain results on how the optimal policy changes with this proportion and traffic intensity. Finally, in Section 6, we present the results of a numerical study we conducted to observe how the performance of the best static policy compares with the generalized $c\mu$ rule under quadratic cost functions and to provide insights into when one should consider such a dynamic policy over the simpler static policies. We summarize our conclusions in Section 7. The proofs of our analytical results are provided in the Appendix. In the following, before we proceed with model description and analysis, we first provide a brief literature review.

2 Literature review

Queueing systems where certain classes of customers have priority over others are called *priority queues*. The study of priority queues dates back to Cobham (1954) who considered a single-server Markovian queueing system (M/M/1) where customers belong to multiple priority classes and the service is non-preemptive, i.e., the service of a customer is not preempted upon arrival of a customer with higher priority. For such a system, Cobham (1954) derived expressions for the long-run average waiting times in the queue for each priority class. This seminal work was followed by Miller (1960) and Jaiswal (1968), who advanced the analysis of priority queues further, e.g., by providing Laplace-Stieltjes transforms of the waiting time distributions for M/G/1 priority queues and considering other priority mechanisms such as preemptive prioritization. When the waiting time of customers is penalized linearly with time, Cox and Smith (1961) established the optimality of the well-known "$c\mu$ rule," which minimizes the long-run average waiting cost in an M/G/1 queue with multiple priority classes. According to the $c\mu$ rule, customers with larger $c_i\mu_i$ index are assigned higher priority, where c_i is the waiting cost per unit time and μ_i is the service rate for type i customers. Following this seminal paper, the optimality of the $c\mu$-type policies has been studied under various settings by Kakalik and Little (1971), Klimov (1974, 1979), Harrison (1975), Pinedo (1983), Nain (1989), Argon and Ziya (2009), Budhiraja et al. (2014) among others, all under the assumption of linear cost functions.

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

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

3 Model description

Consider a single-server queueing system with two types of customers. Customers arrive to the system according to a Poisson process with rate $\lambda > 0$, and an arriving customer belongs to type $i \in \{1, 2\}$ with probability $p_i > 0$, where $p_1+p_2 = 1$, independently of everything else in the system.

Service times for type $i \in \{1, 2\}$ customers are independent and identically distributed (i.i.d.) with mean $\tau_i > 0$ and second moment $\xi_i > 0$. We define $\rho_i \equiv p_i\lambda\tau_i$ and $\rho \equiv \rho_1 + \rho_2$, which we call the system load, and we assume that $\rho < 1$ for stability. Each type i customer incurs a waiting cost $C_i(t)$ when its waiting time in the queue is t, for $t \geq 0$ and $i = 1, 2$. We assume that $C_i(t)$ is first-order differentiable and non-decreasing in t for fixed i.

For such a queueing system, we consider *non-idling* and *non-preemptive* queueing policies that use information only on the type and arrival order of all customers in the system. These policies are non-idling and non-preemptive in the sense that the server does not idle as long as there is a customer in the system and that service of a customer who has been taken into service has to be completed without any preemption before the server moves on to serving another customer. We let Π denote the set of all such queueing policies.

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

$$C_\pi \equiv \lim_{t\to\infty} \frac{\sum_{i=1}^{2} \sum_{k=1}^{n_i(t)} C_i(V_{i,k}^{\pi,x_0})}{t} \tag{1}$$

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

$$C_\pi = \lambda p_1 E\Big[C_1(W_1^\pi)\Big] + \lambda p_2 E\Big[C_2(W_2^\pi)\Big] \tag{2}$$

if Assumption 1 holds for both $i \in \{1, 2\}$:

Assumption 1. *For fixed $i \in \{1, 2\}$ and $\pi \in \Pi$, $E\Big[\big|C_i(W_i^\pi)\big|\Big] < \infty$.*

In Section 5, we show that Assumption 1 holds under some mild conditions when the waiting costs are polynomial.

4 Comparison of FCFS and type-based priority policies under general cost structures

This section presents the main results of the paper, which provide a complete analytical answer to the question we set out to investigate and generalize the classical $c\mu$ rule. The results are somewhat technical in nature and thus their managerial implications may not be readily apparent. The reader should note however that Section 5 builds on these results and provides more insightful results under a variety of conditions.

Specifically, in this section, we compare three policies within Π, namely FCFS, PF_1, and PF_2, which are of special interest due to their common use in practice. To further motivate our focus on these policies, we start with a result that shows that it is sufficient to compare only FCFS, PF_1, and PF_2 to find the optimal policy within Π if $C_i(\cdot)$ is a convex function (in the non-strict sense) for both $i = 1, 2$.

For $i \in \{1, 2\}$, let Π_{P_i} denote the set of non-idling and non-preemptive static policies that prioritize type i customers over type $3 - i$ customers, where the order of service within each type can be based on some specific ordering of arrival times of customers. For example, policies that prioritize type $i \in \{1, 2\}$ customers and apply FCFS or LCFS within each type are in Π_{P_i}. Let also Π_{NP} be the set of all the remaining policies in Π that do not use the type identity of customers in determining the order of service but can use their rank in the queue. For example, FCFS and last-come-last-serve (LCFS) are two policies in this set. Hence, by definition, we have $\Pi = \Pi_{NP} \cup \Pi_{P_1} \cup \Pi_{P_2}$.

Proposition 1. *If $C_1(t)$ and $C_2(t)$ are both convex functions, then*

(a) *$C_F \leq C_\pi$ for any $\pi \in \Pi_{NP}$;*

(b) *$C_{PF_i} \leq C_\pi$ for any $\pi \in \Pi_{P_i}$ and fixed $i \in \{1, 2\}$.*

Proposition 1 implies that when the cost functions for both types are convex, it is sufficient to consider policies in the set $\{F, PF_1, PF_2\}$ instead of the whole set Π. Parts (a) and (b) of Proposition 1 follow directly from Theorem 2 in Vasicek (1977) and Theorem 1 in Haji and Newell (1971), respectively.

4.1 Definitions and lemmas

In order to compare C_F, C_{PF_1}, and C_{PF_2}, we need several definitions and lemmas, which we provide in this subsection.

Definition 1. (E.g., Shaked and Shanthikumar (2007)). Let X and Y be two random variables with corresponding cumulative distribution functions $F_X(\cdot)$ and $F_Y(\cdot)$. If $F_X(x) \geq F_Y(x)$ for all $x \in (-\infty, \infty)$, then X is said to be smaller than Y in the usual stochastic ordering (denoted by $X \leq_{st} Y$).

Definition 2. (Di Crescenzo, 1999). Let X and Y be two non-negative random variables with $X \leq_{st} Y$ and $E[X] < E[Y] < \infty$. Then, we write $Z \equiv \Psi(X, Y)$ to indicate that Z is a random variable with probability density function

$$f_Z(x) = \frac{F_X(x) - F_Y(x)}{E[Y] - E[X]}, x \geq 0, \tag{3}$$

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

$$E[g(Y)] - E[g(X)] = E[g'(Z)]\big(E[Y] - E[X]\big).$$

Lemma 1 presents a probabilistic analogue of the mean value theorem, where Z is a random variable that can be considered as the "mean value" of X and Y. However, unlike for the (deterministic) mean value theorem, Z does not change with the function g, and $Z = \Psi(X, Y)$ is not necessarily ordered (in some stochastic sense) between X and Y. For example, when X and Y are exponential random variables with distinct rates, it can be shown that $Z =_{st} X + Y$ (see Example 3.1 in Di Crescenzo (1999)).

We will use Lemma 1 in several of our results including our main result that compares C_F, C_{PF_1}, and C_{PF_2}. Before we present this result, we need two more lemmas for the comparison of W_i^F, $W_i^{PF_i}$, and $W_{3-i}^{PF_i}$, for $i = 1, 2$.

Lemma 2. *(E.g., Gross et al. (2008) and Miller (1960)) For an M/G/1 queueing system, the expected steady-state waiting times under FCFS and PF_i are given as follows:*

$$E[W^F] = \frac{\lambda\bar{\xi}}{2(1-\rho)}, \quad E[W_i^{PF_i}] = \frac{\lambda\bar{\xi}}{2(1-\rho_i)}, \quad E[W_{3-i}^{PF_i}] = \frac{\lambda\bar{\xi}}{2(1-\rho_i)(1-\rho)},$$

where $\bar{\xi} \equiv p_1\xi_1 + p_2\xi_2$, and we drop the subscript in W_i^F since the distribution of W^F does not depend on i.

Lemma 3. *For fixed $i \in \{1,2\}$, we have $W_i^{PF_i} \leq_{st} W^F \leq_{st} W_{3-i}^{PF_i}$.*

The order of $W_i^{PF_j}$ and W^F for $i,j \in \{1,2\}$, given in Lemma 3 and proved in the Appendix, makes intuitive sense. The steady-state waiting times under FCFS are stochastically less than those for the non-priority type under a type-based priority policy but greater than those for the prioritized type. Lemma 3 specifies a type of stochastic ordering between these three steady-state random variables.

Based on Lemmas 2 and 3, for $i \in \{1,2\}$, we define the following two random variables:

$$U_i^{PF_i} \equiv \Psi(W_i^{PF_i}, W^F) \text{ and } U_{3-i}^{PF_i} \equiv \Psi(W^F, W_{3-i}^{PF_i}).$$

Note that $U_j^{PF_i}$ is well defined for $i,j \in \{1,2\}$ because $W_i^{PF_i} \leq_{st} W^F \leq_{st} W_i^{PF_{3-i}}$ according to Lemma 3, and when $\rho < 1$ and $p_i > 0$, we have $E\big[W_i^{PF_i}\big] < E\big[W^F\big] < E\big[W_i^{PF_{3-i}}\big] < \infty$, for $i \in \{1,2\}$ by Lemma 2.

4.2 Main results

Our main result, namely, Theorem 1, provides necessary and sufficient conditions for the comparison of FCFS, PF_1, and PF_2 under general cost structures.

Theorem 1. *Suppose that Assumption 1 holds for $i \in \{1,2\}$ and $\pi \in \{F, PF_1, PF_2\}$. Then, we have*

(a) $C_F \leq C_{PF_i}$, for $i \in \{1,2\}$, if and only if $a_i \leq b_i$, where

$$a_i \equiv \frac{E\big[C_i'(U_i^{PF_i})\big]}{\tau_i}, \; b_i \equiv \frac{E\big[C_{3-i}'(U_{3-i}^{PF_i})\big]}{\tau_{3-i}}; \text{ and} \tag{4}$$

(b) $C_{PF_1} \leq C_{PF_2}$ if and only if $(1-\rho_1)(a_2 - b_2) \leq (1-\rho_2)(a_1 - b_1)$.

In an immediate corollary to Theorem 1 we provide necessary and sufficient conditions for the optimality of FCFS, PF_1, and PF_2 within the set of these three policies.

Corollary 1. *Suppose that Assumption 1 holds for $i \in \{1,2\}$ and $\pi \in \{F, PF_1, PF_2\}$.*

(a) If $a_1 \leq b_1$ and $a_2 \leq b_2$, then $C_F \leq C_{PF_1}$ and $C_F \leq C_{PF_2}$.

(b) For fixed $i \in \{1,2\}$, if $a_i \geq b_i$ and $(1-\rho_{3-i})(a_i - b_i) \geq (1-\rho_i)(a_{3-i} - b_{3-i})$, then $C_{PF_i} \leq C_F$ and $C_{PF_i} \leq C_{PF_{3-i}}$.

The conditions in Theorem 1 and Corollary 1 require computation of a_i and b_i for $i \in \{1,2\}$. As long as precise expressions for the cost functions $C_i(\cdot)$ are known, it is not difficult to numerically determine a_i and b_i and thus identify which one of the three policies, F, PF_1, and PF_2 performs best. Furthermore under certain assumptions on the structure of the waiting cost functions, it might be possible to come up with closed-form expressions for a_i and b_i or develop precise methods for computing them as we demonstrate for polynomial functions in Section 5.

In order to compute a_i and b_i in Theorem 1 and Corollary 1, we need to obtain $E\left[C_i'(U_i^{PF_j})\right]$ for $i, j \in \{1,2\}$. In certain situations, the cost function may be simple for one type and complicated for the other. For example, it may be the case that the cost function for type 2 customers is linear or quadratic, and the cost function for type 1 customers has a more complex structure. In this case, we can use the next two results to order C_F, C_{PF_1}, and C_{PF_2} without computing $E\left[C_1'(U_1^{PF_j})\right]$ but by computing $E\left[C_2'(U_i^{PF_j})\right]$ for $i, j \in \{1,2\}$.

Corollary 2. *Suppose that Assumption 1 holds for $i \in \{1,2\}$ and $\pi \in \{F, PF_1, PF_2\}$.*

(a) If $C_1'(t) \geq \tau_1 \max\{a_2, b_1\}$ for all $t \geq 0$, then $C_{PF_1} \leq C_F \leq C_{PF_2}$.

(b) If $C_1'(t) \leq \tau_1 \min\{a_2, b_1\}$ for all $t \geq 0$, then $C_{PF_2} \leq C_F \leq C_{PF_1}$.

(c) If $\tau_1 a_2 \leq C_1'(t) \leq \tau_1 b_1$ for all $t \geq 0$, then $C_F \leq C_{PF_1}$ and $C_F \leq C_{PF_2}$.

Corollary 3. *Suppose that Assumption 1 holds for $i \in \{1,2\}$ and $\pi \in \{F, PF_1, PF_2\}$. Assuming $E[C_2'(U_1^{PF_2})] \neq 0$ and $E[C_2'(U_1^{PF_1})] \neq 0$, define*

$$\alpha \equiv \frac{\tau_1 E[C_2'(U_2^{PF_2})]}{\tau_2 E[C_2'(U_1^{PF_2})]} \quad \text{and} \quad \beta \equiv \frac{\tau_1 E[C_2'(U_2^{PF_1})]}{\tau_2 E[C_2'(U_1^{PF_1})]}.$$

(a) If $C_1'(t) \geq \max\{\alpha, \beta\} C_2'(t)$ for all $t \geq 0$, then $C_{PF_1} \leq C_F \leq C_{PF_2}$.

(b) If $C_1'(t) \leq \min\{\alpha, \beta\} C_2'(t)$ for all $t \geq 0$, then $C_{PF_2} \leq C_F \leq C_{PF_1}$.

(c) If $\alpha C_2'(t) \leq C_1'(t) \leq \beta C_2'(t)$ for all $t \geq 0$, then $C_F \leq C_{PF_1}$ and $C_F \leq C_{PF_2}$.

(d) If service times for all customers are i.i.d. and $C_2(\cdot)$ is convex, then $\alpha \leq 1 \leq \beta$.

Corollary 2 compares $C_1'(t)$ with two fixed quantities, namely, $\tau_1 a_2$ and $\tau_1 b_1$, for all $t \geq 0$. Hence, $C_1'(t)$ has to be bounded from either above or below (e.g., when $C_1(t)$ is concave) for the conditions of the corollary to hold. On the other hand, in Corollary 3, we compare $C_1'(t)$ with two time-varying quantities, $\alpha C_2'(t)$ and $\beta C_2'(t)$, and hence $C_1'(t)$ does not need to be bounded. However, in Corollary 3, we require $E\left[C_2'(U_1^{PF_i})\right]$ for $i \in \{1, 2\}$ to be non-zero, which is satisfied when $C_2(\cdot)$ is a strictly increasing function. When $C_2'(t)$ is a constant, i.e., $C_2(t)$ is linear, it can be shown that $\tau_1 a_2 = \tau_1 b_1 = \alpha C_2'(t) = \beta C_2'(t)$, and hence, these two corollaries reduce to one another. We will demonstrate how these two corollaries can be applied when cost functions are polynomial and how they generalize the $c\mu$ rule in Section 5.

5 Comparison of FCFS and type-based priority policies for polynomial cost functions

In this section, we focus on the case where the cost function for at least one type is polynomial. In particular, suppose that for some $i \in \{1, 2\}$,

$$C_i(t) = \sum_{l=1}^{j(i)} h_l(i) t^l, \tag{5}$$

where $j(i) < \infty$ is the degree of the polynomial function $C_i(t)$, and $h_l(i)$ are some real numbers such that $C_i'(t) \geq 0$ for all $t \geq 0$. We first provide conditions under which Assumption 1 holds for type i customers that have a polynomial cost function.

Proposition 2. *For type $i \in \{1, 2\}$ with $C_i(t)$ in the form of (5), if $\rho < 1$ and the first $j(i) + 1$ moments of service times for both types of customers are finite, then $E[(W_i^\pi)^l] < \infty$ for $l = 1, 2, \ldots, j(i)$ and $\pi \in \{F, PF_1, PF_2\}$, and hence, Assumption 1 holds for type i customers under policy $\pi \in \{F, PF_1, PF_2\}$.*

The condition on the moments of service times in Proposition 2 holds for many commonly used distributions such as exponential, gamma, weibull, and lognormal. In the remainder of this section, we assume that this condition holds and hence Assumption 1 holds for customers with polynomial cost functions under $\rho < 1$.

In order to apply Theorem 1 and Corollaries 1, 2, and 3 to the polynomial case, we need to compute $E\left[C_i'(U_k^{PF_m})\right]$ for some $i, k, m \in \{1, 2\}$, where

$$E\left[C_i'(U_k^{PF_m})\right] = \sum_{l=1}^{j(i)} l h_l(i) E\left[\left(U_k^{PF_m}\right)^{l-1}\right]. \tag{6}$$

Here, $E\left[\left(U_k^{PF_m}\right)^{l-1}\right]$ for $l = 1, 2, \ldots, j(i)$ can be computed by the expression

$$E\left[\left(U_k^{PF_m}\right)^{l-1}\right] = \frac{E\left[(W^F)^l\right] - E\left[(W_k^{PF_m})^l\right]}{l\left(E[W^F] - E[W_k^{PF_m}]\right)}, \tag{7}$$

which can be obtained by letting $g(x) = x^l/l$ in Lemma 1. To demonstrate how Theorem 1 and Corollaries 1, 2, and 3 can be used and to gain insights, in the remainder of this section, we focus on polynomial cost functions with a degree of at most two.

5.1 Quadratic cost functions for both customer types

Suppose that $C_i(t) = k_i t^2 + h_i t$, where $k_i, h_i \geq 0$ and $i \in \{1, 2\}$. Let ζ_i denote the third moment of the service times for type $i \in \{1, 2\}$ and $\bar{\zeta} \equiv p_1\zeta_1 + p_2\zeta_2$. Then, Theorem 1 leads to the following proposition that completely characterizes the order of PF_1, PF_2, and FCFS for the case with quadratic cost functions.

Proposition 3. *For quadratic cost functions, the best policy among PF_1, PF_2, and FCFS is characterized as follows: if for some $i = 1, 2$, we have $a_i > b_i$, where*

$$a_i = \frac{k_i}{\tau_i}\left[\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_i \xi_i}{1-\rho_i} + \frac{\xi_{3-i}}{\tau_{3-i}}\right] + \frac{h_i}{\tau_i}, \tag{8}$$

$$b_i = \frac{k_{3-i}}{\tau_{3-i}}\left[\frac{2\bar{\zeta}}{3\bar{\xi}}\left(1 + \frac{1}{1-\rho_i}\right) + \frac{\lambda\bar{\xi}}{1-\rho}\left(1 + \frac{1}{1-\rho_i}\right) + \frac{\xi_i}{\tau_i(1-\rho_i)^2}\right] + \frac{h_{3-i}}{\tau_{3-i}}, \tag{9}$$

then PF_i is the best; and otherwise, i.e., if $a_1 \leq b_1$ and $a_2 \leq b_2$, then FCFS is the best.

Proposition 3 is a generalization of the classical $c\mu$ rule to the quadratic cost setting with possibly the most important difference being that FCFS can now in fact be better than prioritizing either type. To better understand the relation with the $c\mu$ rule, note that Proposition 3 implies that PF_i is the best among PF_1, PF_2, and FCFS, if and only if

$$\frac{k_i}{\tau_i} \geq \frac{\frac{k_{3-i}}{\tau_{3-i}}\left[\frac{2-\rho_i}{1-\rho_i}\left(\frac{2\bar{\zeta}}{3\bar{\xi}}+\frac{\lambda\bar{\xi}}{1-\rho}+\frac{\lambda p_i\xi_i}{1-\rho_i}\right)+\frac{\xi_i}{\tau_i}\right]+\frac{h_{3-i}}{\tau_{3-i}}-\frac{h_i}{\tau_i}}{\left[\frac{2\bar{\zeta}}{3\bar{\xi}}+\frac{\lambda\bar{\xi}}{1-\rho}+\frac{\lambda p_i\xi_i}{1-\rho_i}+\frac{\xi_{3-i}}{\tau_{3-i}}\right]}. \tag{10}$$

One can then recover the $c\mu$ rule by setting $k_1 = k_2 = 0$ in (10).

To gain further insights, we next consider the case where $h_1/\tau_1 = h_2/\tau_2$, e.g., when $C_i(t) = k_i t^2$ for $i \in \{1, 2\}$, in the remainder of this subsection. (Argon et al. (2009) and Ata and Tongarlak (2013) studied similar cost structures.) Then, Proposition 3 leads to the following corollary:

Corollary 4. *For quadratic cost functions, when $h_1/\tau_1 = h_2/\tau_2$ (e.g., when $h_1 = h_2 = 0$), the best policy among PF_1, PF_2, and FCFS is characterized as follows: PF_2 is the best if $k_1/k_2 < A\tau_1/\tau_2$; PF_1 is the best if $k_1/k_2 > B\tau_1/\tau_2$; and FCFS is the best if $A\tau_1/\tau_2 \leq k_1/k_2 \leq B\tau_1/\tau_2$, where*

$$A \equiv \frac{\frac{2\bar{\zeta}}{3\bar{\xi}}+\frac{\lambda\bar{\xi}}{1-\rho}+\frac{\lambda p_2\xi_2}{1-\rho_2}+\frac{\xi_1}{\tau_1}}{\frac{2-\rho_2}{1-\rho_2}\left(\frac{2\bar{\zeta}}{3\bar{\xi}}+\frac{\lambda\bar{\xi}}{1-\rho}+\frac{\lambda p_2\xi_2}{1-\rho_2}\right)+\frac{\xi_2}{\tau_2}} < \frac{\frac{2-\rho_1}{1-\rho_1}\left(\frac{2\bar{\zeta}}{3\bar{\xi}}+\frac{\lambda\bar{\xi}}{1-\rho}+\frac{\lambda p_1\xi_1}{1-\rho_1}\right)+\frac{\xi_1}{\tau_1}}{\frac{2\bar{\zeta}}{3\bar{\xi}}+\frac{\lambda\bar{\xi}}{1-\rho}+\frac{\lambda p_1\xi_1}{1-\rho_1}+\frac{\xi_2}{\tau_2}} \equiv B.$$

Corollary 4 completely characterizes the best policy among FCFS, PF_1, and PF_2 for quadratic cost functions with $h_1/\tau_1 = h_2/\tau_2$ (e.g., when $h_1 = h_2 = 0$). In particular, it states that PF_1 is the best if k_1/k_2 is sufficiently large, PF_2 is the best if k_1/k_2 is sufficiently small, and FCFS is the best if k_1/k_2 is somewhere in between.

Corollary 4 is very useful if the precise values of k_1 and k_2 are known. But what if the service provider has reason to believe that quadratic functions would accurately capture the waiting costs but cannot determine k_1 and k_2 precisely? Of course, in that case, it is impossible to know for sure which policy would be better but one can compare these policies for a range of values for k_1/k_2 (rather than one specific value) and for each policy observe how large the range of k_1/k_2 values that favor that policy over the others is. One would mainly pay attention to plausible values for k_1/k_2 but within those plausible values, the larger the k_1/k_2 interval over which a particular policy is better than the others the more confident one would be for choosing that policy over the others when it comes to implementation. The service provider could also investigate how the k_1/k_2 interval over which each policy is better than the others changes with respect to other system parameters like the arrival rate. Such an analysis would help the provider identify system conditions under which one type of policy would be favored over the others and potentially be more confident about the policy choices particularly when it comes to deciding whether to follow the standard FCFS scheme or go with prioritization with respect to customer types. To that end, we next look at some numerical examples, illustrate how such an analysis can be carried out, and proceed with some analytical results that provide support to some of our observations.

Figure 1 provides plots for four numerical examples that demonstrate how the best policy can be determined using Corollary 4 and how the regions of optimality change with ρ (or equivalently with λ). From the figure, we can observe that the region where FCFS is the best policy enlarges as λ increases. This suggests that given the uncertainty around the true value of k_1/k_2, higher arrival rates make it increasingly more likely for type-based priority policies to perform worse than the standard FCFS policy. One should note however that while we can observe that both A and B monotonically change with λ, they are not necessarily increasing or decreasing in all cases. We next prove this monotonicity property and provide necessary and sufficient conditions under which A and B increase or decrease with respect to λ.

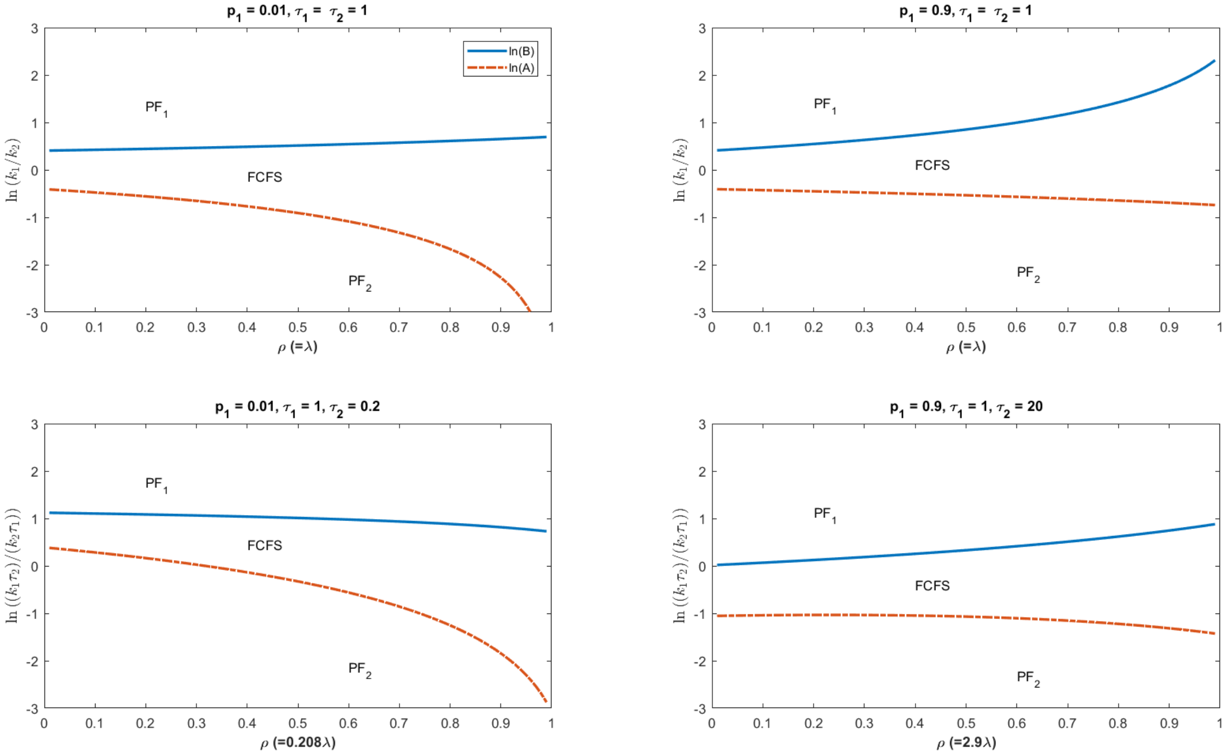

Figure 1: Regions of optimality for FCFS, PF_1, and PF_2 as a function of ρ (or equivalently λ) under quadratic waiting costs with $h_1/\tau_1 = h_2/\tau_2$ and exponential service times.

Proposition 4. *(a) A decreases in λ if and only if*

$$\frac{\xi_2}{\tau_2} - \frac{(2-\rho_2)\xi_1}{(1-\rho_2)\tau_1} < \frac{\frac{p_2\tau_2}{(1-\rho_2)^2}\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_2\xi_2}{1-\rho_2}\right)\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_2\xi_2}{1-\rho_2} + \frac{\xi_1}{\tau_1}\right)}{\frac{\bar{\xi}}{(1-\rho)^2} + \frac{p_2\xi_2}{(1-\rho_2)^2}}. \tag{11}$$

(b) B increases in λ if and only if

$$\frac{\xi_1}{\tau_1} - \frac{(2-\rho_1)\xi_2}{(1-\rho_1)\tau_2} < \frac{\frac{p_1\tau_1}{(1-\rho_1)^2}\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_1\xi_1}{1-\rho_1}\right)\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_1\xi_1}{1-\rho_1} + \frac{\xi_2}{\tau_2}\right)}{\frac{\bar{\xi}}{(1-\rho)^2} + \frac{p_1\xi_1}{(1-\rho_1)^2}}. \tag{12}$$

(c) As $\lambda \to 1/\bar{\tau}$, where $\bar{\tau} \equiv p_1\tau_1 + p_2\tau_2$, we have $A \to \frac{p_1\tau_1}{2p_1\tau_1+p_2\tau_2}$ and $B \to \frac{p_1\tau_1+2p_2\tau_2}{p_2\tau_2}$.

Parts (a) and (b) of Proposition 4 provide necessary and sufficient conditions under which the thresholds for the optimality of the three policies (see Corollary 4) monotonically change with λ. One can obtain a simpler sufficient condition by noting that the right-hand sides of (11) and (12) are both nonnegative: if $\frac{\xi_1}{\tau_1} > \frac{\xi_2(1-\rho_2)}{\tau_2(2-\rho_2)}$ (which holds if $\frac{\xi_1}{\tau_1}/\frac{\xi_2}{\tau_2} > 1/2$), then A decreases in λ and if $\frac{\xi_1}{\tau_1} < \frac{(2-\rho_1)\xi_2}{(1-\rho_1)\tau_2}$ (which holds if $\frac{\xi_1}{\tau_1}/\frac{\xi_2}{\tau_2} < 2$), then B increases in λ. This means that if ξ_i/τ_i values for the two customer types are relatively close to each other, specifically, one is not more than twice as large as the other, then the region where FCFS is the best gets larger while the regions for the type-based priority policies get smaller suggesting that higher arrival rates increasingly favor FCFS over the prioritization policies.

Note also that as $p_2 \to 0$, the opposite of (11) will hold if and only if $\frac{\xi_1}{\tau_1}/\frac{\xi_2}{\tau_2} < 1/2$, in which case PF_2 is preferred for a larger range of values of k_1/k_2 as λ increases. Similarly, from (12), we find that B decreases in λ when $p_1 \to 0$ and $\frac{\xi_1}{\tau_1}/\frac{\xi_2}{\tau_2} > 2$, and thus PF_1 is preferred for a larger range of values of k_1/k_2 as λ increases. Thus, we can conclude that if the proportion of one type of customers is sufficiently small, but the ratio ξ_i/τ_i for the same type is sufficiently large (at least twice as large), then prioritizing that type becomes more preferable under a larger set of k_1/k_2 values and thus more likely to be the right choice as λ increases.

When interpreting these findings, it would be useful to note what the ratio ξ_i/τ_i represents. For example, for fixed τ_i's, a higher ξ_i would imply a higher variance. Hence, if the mean and variance for service times are similar for the two types, then higher arrival rates increasingly favor FCFS over type-based priority policies. On the other hand, if the mean service time for the two types are similar but the variance is much higher for one of the types, then higher arrival rates will increasingly favor giving priority to the type with higher variance if the proportion of that type is sufficiently small.

Proposition 4(c) provides the limiting values of A and B under a heavy traffic condition and thus can be used to precisely describe the regions under which each one of the three policies would perform better than the others when the system is heavily loaded. Noting the fact that A converges to a value that is less than 1 and B converges to a value that is larger than 1, we can also conclude that FCFS should be preferred in heavy traffic regardless of the service time distribution of either type if k_1/τ_1 and k_2/τ_2 are similar.

We next study the effects of p_1 (and hence p_2) on the comparison of FCFS, PF_1 and PF_2. We first provide four numerical examples in Figure 2. We notice from Figure 2 that A and B do not change monotonically in p_1 except when the service times for all customers are i.i.d. In our next result, we prove monotonicity of A and B in p_1 under i.i.d. service times, and also provide the limiting values of A and B in heavy traffic and as p_1 approaches 0 or 1.

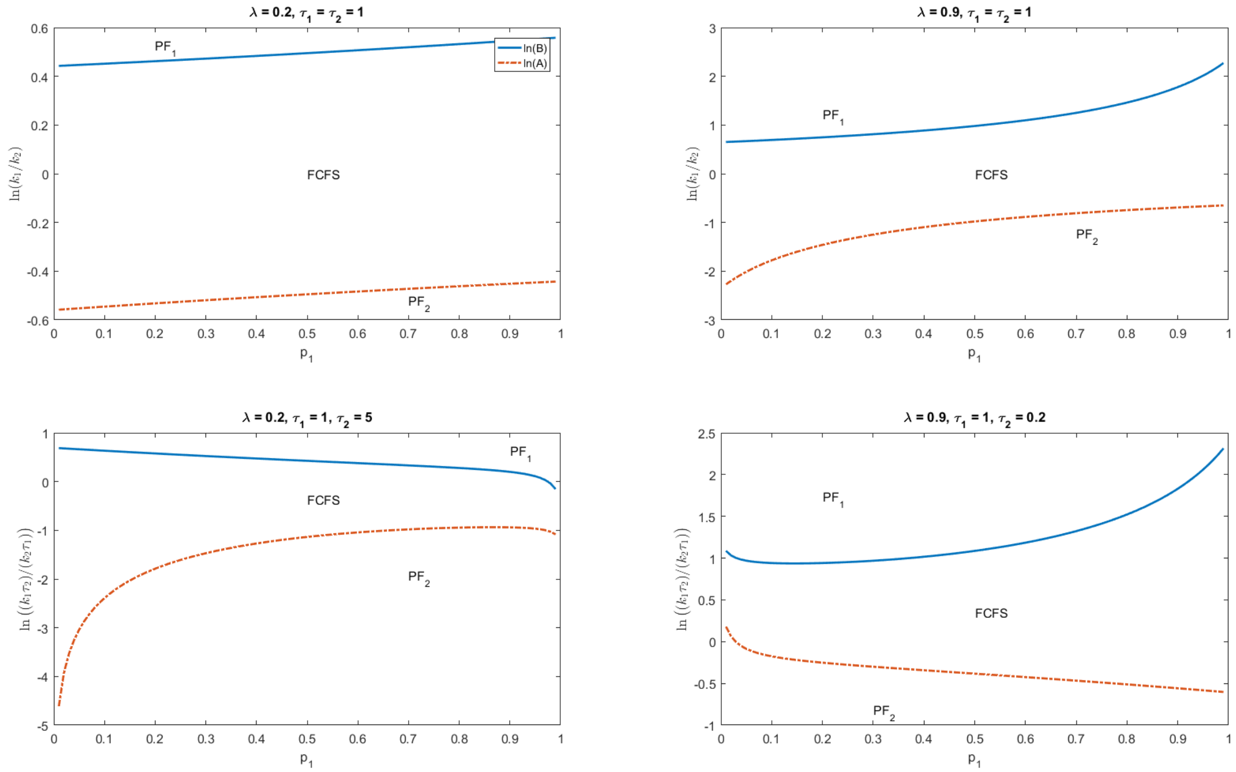

Figure 2: Regions of optimality for FCFS, PF_1, and PF_2 as a function of p_1 under quadratic waiting costs with $h_1/\tau_1 = h_2/\tau_2$ and exponential service times.

Proposition 5. *(a) When service times are i.i.d. for all customers, A and B both increase in p_1 (and hence decrease in p_2).*

(b) $\lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 0} A = 0, \quad \lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 0} B = 2, \quad \lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 1} A = \frac{1}{2}, \quad \lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 1} B = \infty.$

Proposition 5(a) indicates that if service times are i.i.d., as the proportion of one type increases, giving priority to that type becomes preferable for a smaller range of k_1/k_2 values while prioritizing the other type is preferred for a wider range (see also the top two plots in Figure 2). This is because, unlike the case where waiting costs are linear, under quadratic waiting costs, long waits are penalized very heavily and as the proportion of higher priority type increases, the waiting cost incurred by the lower priority customers increases significantly. (Nevertheless, as we can see from the bottom two plots in Figure 2, this intuition does not work when service times are not identically distributed for all customers since in that case the type given priority also influences the rate at which waiting customers progress in the queue.) Similarly, Proposition 5(b) implies that under heavy traffic, type i customers should not be prioritized if the proportion of this type is close to one; instead, the other type, i.e., type $3-i$, should be served first if $k_{3-i}/\tau_{3-i} > 2k_i/\tau_i$, and otherwise FCFS should be applied. In other words, when the proportion of one customer significantly dominates the other, under heavy traffic, giving priority can only be justified for the class with the small proportion and that justification requires that its k_i/τ_i is at least twice as large as that of the other type. Otherwise, it is better to use FCFS.

Finally, we compare the values of A and B under two different service time distributions with the same means. Let $A_{exp}[B_{exp}]$ and $A_{det}[B_{det}]$ denote the values of $A[B]$ under exponential and deterministic service times, respectively.

Proposition 6. *(a)* $A_{exp} \leq A_{det}$ *if and only if* $\tau_2 \leq \tau_1(2-\rho_2)/(1-\rho_2)$.

(b) $B_{exp} \geq B_{det}$ *if and only if* $\tau_2 \geq \tau_1(1-\rho_1)/(2-\rho_1)$.

Proposition 6 implies that if $\frac{\tau_1}{\tau_2} \in \left(\frac{1-\rho_2}{2-\rho_2}, \frac{2-\rho_1}{1-\rho_1}\right)$, then $A_{exp} \leq A_{det} < B_{det} \leq B_{exp}$, and hence, when the mean service times are not significantly different for the two types, FCFS is preferable for a wider range of values of k_1/k_2 under exponential service times than under deterministic service times. This suggests that when the two types are not too different in terms of mean service times, higher service time variability makes FCFS a better choice under a larger range of waiting cost scenarios. When service times have higher variance, waiting times will also have higher variance regardless of whether FCFS or a type-based priority policy is in place. Nevertheless, due to the convexity of the waiting cost functions, the impact will be larger on the type-based 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, $\tau_1/\tau_2 > (2-\rho_1)/(1-\rho_1)$, then $A_{exp} \leq A_{det}$ and $B_{exp} \leq B_{det}$, which implies that under deterministic service times, PF_2 (prioritizing the faster type) is preferred for a wider range of values of k_1/k_2, and PF_1 (prioritizing the slower type) is preferred for a narrower range of values of k_1/k_2 than that under exponential service times.

5.2 Minimizing the variance of waiting times in steady state

In this section, we discuss how the results from Section 5.1 can be used to derive insights into the problem of minimizing the variance of the 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 interest especially in the context of fairness in queueing systems. In particular, Kingman (1962), Avi-Itzhak and Levy (2004), and references therein use variance of waiting times as a measure of fairness in a queueing system in that a policy that has a smaller variance of waiting times is regarded as a fairer policy. Kingman (1962) and Vasicek (1977) prove that FCFS minimizes the variance of waiting times among all non-idling queueing disciplines and thus is the "fairest" discipline for various queueing systems but under the assumption that customers are indistinguishable, i.e., there is a single class of customers. Avi-Itzhak and Levy (2004) propose a new fairness measure that computes the expected number of positions that a job is pushed ahead or backwards under a policy compared to FCFS 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, unlike this paper, all earlier work on minimization of variance of waiting times considered customers belonging to a single class. We next use our results on quadratic cost functions to study the variance minimization problem for an M/G/1 queue with two classes of customers with equal mean service times but distinct service-time distributions.

For identical mean service times for all customers, i.e., $\tau_1 = \tau_2$, the steady-state mean waiting times are the same under FCFS, PF_1, and PF_2, as can be verified using Lemma 2. Hence, minimizing the variance of the steady-state waiting times within Π is equivalent to minimizing the second moment of the steady-state waiting times, which corresponds to letting $C_1(t) = C_2(t) = t^2$ for $t \geq 0$ in our formulation. Corollary 4 then immediately yields the following result.

Corollary 5. *When all customers have equal mean service times, the variance of the steady-state waiting times among all policies in* Π *is minimized by* PF_2 *if*

$$(p_1\xi_1 + p_2\xi_2)\left[\Big(1 - \frac{\rho(1-\rho_2)}{1-\rho}\Big)\xi_1 - \Big(1 - \rho_2 + \frac{\rho_2}{1-\rho} + \frac{\rho_2}{1-\rho_2}\Big)\xi_2\right] > \frac{2\bar{\tau}\bar{\zeta}}{3}; \tag{13}$$

by PF_1 *if*

$$(p_1\xi_1 + p_2\xi_2)\left[\Big(1 - \frac{\rho(1-\rho_1)}{1-\rho}\Big)\xi_2 - \Big(1 - \rho_1 + \frac{\rho_1}{1-\rho} + \frac{\rho_1}{1-\rho_1}\Big)\xi_1\right] > \frac{2\bar{\tau}\bar{\zeta}}{3}; \tag{14}$$

and by FCFS otherwise.

Corollary 5 appears to be somewhat technical at first but a closer examination of Conditions (13) and (14) after some algebraic manipulations leads to some interesting insights into the problem of minimizing the steady-state waiting time variance:

- When $\rho \geq \sqrt{2}/(1+\sqrt{2})$ and $(1-\rho)^2/\rho^2 \leq p_1 \leq 1 - (1-\rho)^2/\rho^2$, the left-hand sides of both (13) and (14) are non-positive, and hence, regardless of the service time distributions, FCFS provides the smallest variance for the steady-state waiting times. In other words, when the traffic intensity is sufficiently large and neither type is dominant in numbers, then FCFS is better than all other static policies. Furthermore, as ρ increases, the need for balance between p_1 and p_2 for FCFS to be better than the other two policies diminishes and becomes completely unnecessary as ρ approaches one. This is consistent with the asymptotic optimality of the generalized $c\mu$ rule, which reduces to FCFS when mean service times are the same for both types and the cost functions are given by $C_1(t) = C_2(t) = t^2$ for $t \geq 0$.

- When $\rho \geq \sqrt{2}/(1+\sqrt{2})$ and $p_i \leq (1-\rho)^2/\rho^2$ for some type i, then PF_{3-i} [FCFS] is the best static policy if the service-time variance of type $3-i$ is sufficiently small [large]. In other words, when the traffic intensity and the proportion of one type are sufficiently large, then prioritizing the type with a larger proportion of demand is the best if its service time variance is small enough; otherwise, it is best to use FCFS.

- When $\rho < \sqrt{2}/(1+\sqrt{2})$, then the type with a sufficiently smaller service-time variance should be prioritized. If the service time variances are not too different (e.g., $\xi_1 = \xi_2$), then FCFS becomes the best static policy even if the traffic intensity is not large. This generalizes the earlier work by Kingman (1962) and Vasicek (1977) that showed that the variance for waiting times is minimized by the FCFS policy under i.i.d. service times for all customers.

5.3 Quadratic cost for one type and general cost for the other type

Suppose one type of customers has a quadratic cost function (say, $C_2(t) = k_2t^2 + h_2t$ for $h_2, k_2 \geq 0$), and the other type has a general cost function that is not necessarily in quadratic form. In this section, we will demonstrate how Corollaries 2 and 3 can be used for this case assuming that Assumption 1 holds for both types.

We first focus on Corollary 2. When $C_2(t)$ is a quadratic function, a_2 and b_1 are given by (8) and (9), respectively, and hence we have $a_2 < b_1$ (see the proof of Proposition 3 in the Appendix). Therefore, in Corollary 2, we can replace $\max\{a_2, b_1\}$ with b_1 and $\min\{a_2, b_1\}$ with a_2. Furthermore, since $a_2 < b_1$, we know that the interval $(\tau_1 a_2, \tau_1 b_1)$ is not necessarily an empty set and hence part (c) of Corollary 2 could be applicable. Consequently, Corollary 2 implies that if the smallest value the derivative of $C_1(t)$ takes is at least $\tau_1 b_1$, then type 1 customers should be prioritized; if the largest value the derivative of $C_1(t)$ takes is at most $\tau_1 a_2$, then type 2 customers should be prioritized; and if the derivative of $C_1(t)$ lies between $\tau_1 a_2$ and $\tau_1 b_1$ at all times, then FCFS should be employed. Furthermore, by Equations (8) and (9), we notice that a_i, b_i and the difference $b_{3-i} - a_i$ all increase in λ for $i \in \{1, 2\}$ since ρ_i, ρ_{3-i}, $1/(1-\rho)$, $1/(1-\rho_i)$ and $1/(1-\rho_{3-i})$ all increase in λ. This implies that the bounds $\tau_1 a_2$ and $\tau_1 b_1$, and the length of the interval $(\tau_1 a_2, \tau_1 b_1)$ are all increasing as λ becomes larger. Moreover, both a_2 and b_1 go to infinity as λ approaches $\bar{\tau}^{-1}$. Combining this with Corollary 2 leads to an important conclusion: if the derivative of the cost function for one type is bounded from above and the other type has a quadratic cost function, then it is best to prioritize the type with quadratic cost under heavy traffic no matter what the service time and cost parameters are. To better understand these implications and where they can be useful, it will be helpful to consider a few examples:

Example 1. (i) If $C_1(t) = h_1 t$ for $t \geq 0$, where h_1 is a finite and positive constant, then $C_1'(t) = h_1$ is bounded. Hence, Corollary 2 leads to a complete characterization of the best policy among FCFS, PF_1, and PF_2 in this case: PF_1 is preferred if $h_1 \geq \tau_1 b_1$, PF_2 is preferred if $h_1 \leq \tau_1 a_2$, and FCFS is preferred otherwise. Note also that as λ increases, the range of h_1 values where PF_1 [PF_2] is preferred shrinks [enlarges] and the range for which FCFS is preferred shifts up and becomes wider. Furthermore, since $a_2 \to \infty$ as $\lambda \to 1/\bar{\tau}$, PF_2 is preferred for any finite h_1 under heavy traffic. This means that when type 1 customers have linear and type 2 customers have quadratic waiting costs, prioritizing type 2 customers will reduce the long-run average cost in heavy traffic no matter what the cost and service time parameters are.

(ii) If $C_1(t) = h_1 \ln(t+1)$ for $t \geq 0$ and positive constant h_1, we have $C_1'(t) \leq h_1$ for all $t \geq 0$, and hence C_{PF_2} is the smallest if $h_1 \leq \tau_1 a_2$. As $\lambda \to 1/\bar{\tau}$, the bound $\tau_1 a_2$ goes to infinity, which indicates that PF_2 is the best for any h_1 under heavy traffic.

(iii) If $C_1(t) = k_1(e^{h_1 t} - 1)$ for $t \geq 0$ and positive constants k_1 and h_1, we have $C_1'(t) \geq k_1 h_1$ for all $t \geq 0$, and hence C_{PF_1} is the smallest if $k_1 h_1 \geq \tau_1 b_1$. As $\lambda \to 1/\bar{\tau}$, the bound $\tau_1 b_1$ goes to infinity, and hence in this case we do not obtain a sufficient condition for PF_1 to be the best policy.

We next consider Corollary 3 under the case where the waiting cost for type 2 customers is a quadratic function. To demonstrate how Corollary 3 could be used, we first applied it to functions given in Example 1 and also identified its differences from Corollary 2. In particular, we showed that both Corollaries 2 and 3 could be useful in different situations. The interested reader is referred to Example 3 in the Appendix. We then obtained the following result that tells us more about how the optimality regions for the three policies under comparison change in the case where the cost for one type is quadratic but the other is general.

Proposition 7. *When $C_2(t)$ is a quadratic function, α and β in Corollary 3 are given by*

$$\begin{aligned} \alpha &= \left(\frac{\tau_1}{\tau_2}\right) \frac{k_2\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}(2-\rho-\rho_2)}{(1-\rho)(1-\rho_2)} + \frac{\lambda p_1 \xi_1 (1-\rho)}{\rho_1(1-\rho_2)}\right) + h_2}{k_2\left(\frac{2\bar{\zeta}(2-\rho_2)}{3\bar{\xi}(1-\rho_2)} + \frac{\lambda\bar{\xi}(2-\rho_2)}{(1-\rho)(1-\rho_2)} + \frac{\lambda p_2 \xi_2}{\rho_2(1-\rho_2)^2}\right) + h_2}, \\ \beta &= \left(\frac{\tau_1}{\tau_2}\right) \frac{k_2\left(\frac{2\bar{\zeta}(2-\rho_1)}{3\bar{\xi}(1-\rho_1)} + \frac{\lambda\bar{\xi}(2-\rho_1)}{(1-\rho)(1-\rho_1)} + \frac{\lambda p_1 \xi_1}{\rho_1(1-\rho_1)^2}\right) + h_2}{k_2\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}(2-\rho-\rho_1)}{(1-\rho)(1-\rho_1)} + \frac{\lambda p_2 \xi_2 (1-\rho)}{\rho_2(1-\rho_1)}\right) + h_2}. \end{aligned}$$

Furthermore, we have the following:

(a) $\alpha < \beta$.

(b) If (11) *holds, then α decreases in λ, and if* (12) *holds, then β increases in λ. (When $h_2 = 0$, conditions* (11) *and* (12) *are also necessary for the respective results.) Moreover,*

$$\lim_{\lambda \to 1/\bar{\tau}} \alpha = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{p_1\tau_1}{2p_1\tau_1 + p_2\tau_2}\right), \quad \lim_{\lambda \to 1/\bar{\tau}} \beta = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{p_1\tau_1 + 2p_2\tau_2}{p_2\tau_2}\right).$$

(c) When service times are i.i.d. for all customers, α and β both increase in p_1 (and hence decrease in p_2).

By Proposition 7(a), we can replace $\max\{\alpha, \beta\}$ with β and $\min\{\alpha, \beta\}$ with α in Corollary 3. In addition, when service times are i.i.d. (so that both (11) and (12) hold), Proposition 7(b) implies that as λ increases α decreases and β increases which means that when the system becomes more congested, the region where FCFS is preferred becomes larger. This observation is in agreement with our conclusions for the case where waiting cost functions for both customer types are quadratic, and thus strengthens the idea that FCFS becomes increasingly more favorable with higher arrival rates under a large class of waiting cost functions.

5.4 Linear cost for one type and general cost for the other type

Suppose that one type of customers has a linear cost function (say, $C_2(t) = h_2 t$ for $t \geq 0$ and $h_2 > 0$). Then, we have $a_2 = b_1 = h_2/\tau_2$ and $\alpha = \beta = \tau_1/\tau_2$, which reduces Corollaries 2 and 3 to the same result:

(a) If $C_1'(t) \geq \frac{h_2\tau_1}{\tau_2}$ for all $t \geq 0$, then $C_{PF_1} \leq C_F \leq C_{PF_2}$.

(b) If $C_1'(t) \leq \frac{h_2\tau_1}{\tau_2}$ for all $t \geq 0$, then $C_{PF_2} \leq C_F \leq C_{PF_1}$.

We next discuss what this result implies for functions given in Example 1. For notational simplicity, let $\mu_i = 1/\tau_i$ for $i \in \{1, 2\}$.

Example 2. (i) When $C_1(t) = h_1 t$ for $t \geq 0$ and positive constant h_1, PF_1 is preferred if $h_1\mu_1 \geq h_2\mu_2$ and PF_2 is preferred otherwise. This is the well-known $c\mu$ rule, which indicates that under linear cost functions we should give priority to the type with the larger $c\mu$ value (in our notation, the larger $h\mu$ value), see, e.g., Cox and Smith (1961).

(ii) When $C_1(t) = h_1 \ln(t+1)$ for $t \geq 0$ and positive constant h_1, PF_2 is the best if $h_1\mu_1 \leq h_2\mu_2$.

(iii) When $C_1(t) = k_1(e^{h_1 t} - 1)$ for $t \geq 0$ and positive constants k_1 and h_1, PF_1 is the best if $k_1 h_1 \mu_1 \geq h_2\mu_2$.

Based on Example 2, one might conjecture that FCFS cannot be the best policy if the cost function for one type is linear. However, this is not true as we have seen in Example 1(i) that FCFS can be better than the type-based priority policies if the waiting cost for one type is linear and that for the other type is quadratic (since a_2 is strictly less than b_1 by (27)).

6 Numerical study

The main objective of this section is to investigate how the performances of state-independent policies FCFS, PF_1, and PF_2 compare with that of the state-dependent, more complex alternative, namely the generalized $c\mu$ (G-$c\mu$) rule. In particular, we aim to identify conditions under which it may be worthwhile to use the G-$c\mu$ rule as opposed to the simpler alternatives and also the conditions under which the additional complexity of the G-$c\mu$ rule does not seem to bring much benefit. Although G-$c\mu$ rule is not necessarily an optimal dynamic policy, it is shown to be asymptotically optimal for convex cost functions under heavy traffic (Van Mieghem, 1995).

In our experimental setup, we considered cost functions of the form $C_1(t) = kt^2$ and $C_2(t) = t^2$, $t \geq 0$, for different values of $k > 0$. Service times for type $i \in \{1, 2\}$ customers are exponentially distributed with mean τ_i, where τ_2 is fixed at one unit of time. We consider 81 different scenarios corresponding to all combinations of $\rho \in \{0.3, 0.7, 0.9\}$, $p_1 \in \{0.1, 0.5, 0.9\}$, $\tau_1 \in \{0.2, 1, 5\}$, and $k \in \{0.1, 0.9, 5\}$. In order to find the optimal static policy within Π (denoted by π^*) for these scenarios, we computed the corresponding values of A and B using Corollary 4 as reported in Table 1. Recall that by Corollary 4, PF_2 has the smallest cost if $k < A\tau_1$; PF_1 has the smallest cost if $k > B\tau_1$; and FCFS has the smallest cost if $A\tau_1 \leq k \leq B\tau_1$. We then were able to compute the long-run average cost under the optimal policy within Π (denoted by C_{π^*}) using analytical expressions for C_F, C_{PF_1}, or C_{PF_2}.

Table 1: The threshold values to characterize the best policy in Π.

		$\tau_1 = 0.2$		$\tau_1 = 1$		$\tau_1 = 5$	
ρ	p_1	$A\tau_1$	$B\tau_1$	$A\tau_1$	$B\tau_1$	$A\tau_1$	$B\tau_1$
0.3	0.1	0.075	0.252	0.532	1.612	4.025	12.79
	0.5	0.076	0.253	0.580	1.725	3.955	13.09
	0.9	0.078	0.248	0.620	1.880	3.970	13.35
0.7	0.1	0.048	0.318	0.307	1.831	2.540	12.76
	0.5	0.058	0.330	0.450	2.223	3.030	17.35
	0.9	0.078	0.394	0.546	3.255	3.140	21.01
0.9	0.1	0.021	0.370	0.169	2.000	1.735	12.80
	0.5	0.040	0.395	0.375	2.664	2.530	25.00
	0.9	0.078	0.576	0.500	5.918	2.700	46.56

To obtain the long-run average cost under the G-$c\mu$ rule (denoted by C_G), we simulated the underlying queueing system on Arena 14 simulation software. Specifically, we ran 100 independent replications of length 60,000 minutes for each scenario and truncated the first 6,000 minutes based on a warm-up period analysis. To implement the G-$c\mu$ rule in our simulations, we computed a priority index for each customer in the queue and assigned non-preemptive priority to the one with the largest index. Under the specific cost structure and experimental setting of this section, the priority index for a customer who waited for $t \geq 0$ time units is given by $2kt/\tau_1$ for a type 1 customer and $2t$ for a type 2 customer. We report the mean percentage change in cost by using G-$c\mu$ rule over the best static policy, i.e., $(C_G - C_{\pi^*}) \times 100/C_{\pi^*}$ and the 95% confidence interval (C.I.) of this percentage change from the simulation runs in Tables 2, 3, and 4. If this confidence interval does not contain zero, then we conclude that there is statistical evidence that the best static policy and the G-$c\mu$ rule are different and the comparison is in favor of the best static policy for a positive confidence interval and the G-$c\mu$ rule for a negative confidence interval. (Here, and in the rest of this section "the best static policy" or "the optimal static policy" both refer to the optimal policy in Π.)

We first present the case with equal service rates for all customers in Table 2. From Table 2, we find that the differences between the best static policy and the G-$c\mu$ rule are not statistically significant in most scenarios. (Confidence intervals implying statistical significance are indicated in bold.) In particular, for the case with equal service rates for all customers, when the waiting costs are not too different ($k = 0.9$) or the traffic intensity is not high ($\rho \in \{0.3, 0.7\}$), there seems to be no advantage to using the G-$c\mu$ rule over the best static policy. Indeed, in two scenarios ($\rho = 0.3, p_1 = 0.9, k = 0.1$ and $\rho = 0.7$, $p_1 = 0.5$, $k = 0.1$), the best static policy performs better than the G-$c\mu$ rule. However, when the traffic intensity is high ($\rho = 0.9$), the difference in costs between the two types is large ($k \in \{0.1, 5\}$), and the proportion of the "important" type with a higher cost coefficient is large, then the G-$c\mu$ rule performs significantly better than the best static policy.

We next compare the best static policy and the G-$c\mu$ rule under different service rates in Tables 3 and 4. Below are our observations:

- Under light traffic, there is no statistically significant difference between the best static policy and the G-$c\mu$ rule in most scenarios, and in statistically significant ones, the best static policy performs better.

- For moderate or high traffic intensity, when there is a clearly more important type that has a

Table 2: The best static policy (π^*) and 95% C.I. on the percentage change in cost by using G-$c\mu$ rule over the best static policy when $\tau_1 = \tau_2 = 1$.

		$k = 0.1$		$k = 0.9$		$k = 5$	
ρ	p_1	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$
0.3	0.1	PF_2	-0.38 ± 1.60	FCFS	-1.32 ± 1.60	PF_1	-0.62 ± 1.58
	0.5	PF_2	0.98 ± 1.60	FCFS	-1.29 ± 1.60	PF_1	0.32 ± 1.56
	0.9	PF_2	**1.67 ± 1.60**	FCFS	-1.17 ± 1.60	PF_1	-0.97 ± 1.61
0.7	0.1	PF_2	1.41 ± 1.51	FCFS	-0.06 ± 1.67	PF_1	0.13 ± 1.63
	0.5	PF_2	**3.37 ± 1.38**	FCFS	-0.25 ± 1.66	PF_1	-0.57 ± 1.50
	0.9	PF_2	0.99 ± 1.58	FCFS	-0.11 ± 1.66	PF_1	**-2.92 ± 1.57**
0.9	0.1	PF_2	**-7.52 ± 4.53**	FCFS	2.60 ± 5.39	PF_1	1.39 ± 5.31
	0.5	PF_2	1.01 ± 5.00	FCFS	2.42 ± 5.38	PF_1	**-6.87 ± 4.85**
	0.9	PF_2	2.58 ± 5.35	FCFS	2.59 ± 5.39	FCFS	**-18.50 ± 4.19**

substantially higher cost parameter, service rate, and proportion of demand (scenarios with $\tau_1 = 5, k = 0.1, p_1 = 0.1$ or $\tau_1 = 0.2, k = 5, p_1 = 0.9$), then the optimal static policy is preferable over the G-$c\mu$ rule.

- Under heavy traffic, when the two types are similar in terms of their cost parameters ($k = 0.9$) but there is a substantial difference between their service rates and the faster type has a higher proportion (scenarios with $\tau_1 = 0.2, p_1 = 0.9$ or $\tau_1 = 5, p_1 = 0.1$), then the G-$c\mu$ rule outperforms the optimal static policy.

Table 3: The best static policy (π^*) and 95% C.I. on the percentage change in cost by using G-$c\mu$ rule over the best static policy when $\tau_1 = 5$ and $\tau_2 = 1$.

		$k = 0.1$		$k = 0.9$		$k = 5$	
ρ	p_1	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$
0.3	0.1	PF_2	-0.65 ± 2.95	PF_2	2.14 ± 3.29	FCFS	-0.49 ± 3.70
	0.5	PF_2	0.52 ± 2.22	PF_2	1.11 ± 2.50	FCFS	0.00 ± 3.11
	0.9	PF_2	0.00 ± 2.78	PF_2	0.56 ± 3.00	FCFS	0.31 ± 3.12
0.7	0.1	PF_2	**1.90 ± 1.73**	PF_2	2.87 ± 3.23	FCFS	-2.89 ± 3.74
	0.5	PF_2	**3.12 ± 2.62**	PF_2	2.46 ± 4.55	FCFS	1.48 ± 4.56
	0.9	PF_2	1.26 ± 4.67	PF_2	1.76 ± 4.88	FCFS	1.75 ± 5.17
0.9	0.1	PF_2	**6.27 ± 6.03**	PF_2	**-12.13 ± 7.44**	FCFS	-1.41 ± 8.87
	0.5	PF_2	0.45 ± 9.18	PF_2	1.41 ± 9.97	FCFS	0.31 ± 9.72
	0.9	PF_2	1.82 ± 9.03	PF_2	-0.17 ± 8.51	FCFS	0.36 ± 8.79

Another important observation from the numerical experiments presented in Tables 2, 3, and 4

Table 4: The best static policy (π^*) and 95% C.I. on the percentage change in cost by using G-$c\mu$ rule over the best static policy when $\tau_1 = 0.2$ and $\tau_2 = 1$.

		$k = 0.1$		$k = 0.9$		$k = 5$	
ρ	p_1	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$	π^*	$\frac{C_G - C_{\pi^*}}{C_{\pi^*}} \times 100$
0.3	0.1	FCFS	-1.24 ± 1.59	PF_1	-0.70 ± 1.60	PF_1	-0.56 ± 1.49
	0.5	FCFS	-1.23 ± 1.47	PF_1	**1.70 ± 1.28**	PF_1	1.06 ± 1.08
	0.9	FCFS	-1.04 ± 1.72	PF_1	**2.50 ± 1.13**	PF_1	0.84 ± 1.08
0.7	0.1	FCFS	0.01 ± 1.73	PF_1	0.58 ± 1.81	PF_1	1.63 ± 1.73
	0.5	FCFS	**-2.48 ± 1.77**	PF_1	1.73 ± 1.87	PF_1	**2.84 ± 1.42**
	0.9	FCFS	**-4.05 ± 2.05**	PF_1	**3.19 ± 1.43**	PF_1	**5.14 ± 0.99**
0.9	0.1	FCFS	1.91 ± 5.60	PF_1	3.30 ± 5.60	PF_1	3.65 ± 5.29
	0.5	FCFS	-1.28 ± 4.40	PF_1	0.98 ± 4.63	PF_1	1.28 ± 4.88
	0.9	FCFS	**-4.78 ± 3.95**	PF_1	**-11.02 ± 3.34**	PF_1	**7.00 ± 2.61**

is that when FCFS is the best static policy, it either performs similarly with the G-$c\mu$ rule or the G-$c\mu$ rule outperforms it. We also observe that for heavy-traffic scenarios where the parameters fall close to the thresholds that characterize the optimal static policy reported in Table 1, possibly suggesting that none of the static policies stands out, the G-$c\mu$ rule performs better than the optimal static policy. Hence, it would be worthwhile to consider the more complex G-$c\mu$ rule over a static 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 optimal static policy. If they are closer to a threshold, such as in scenarios $\tau_1 = 1, k = 5, p_1 = 0.9$ or $\tau_1 = 5, k = 0.9, p_1 = 0.1$ above, then this could be taken as an indicator that there is not a clearly more important type and hence G-$c\mu$ rule should be considered under heavy traffic. On the other hand, when the traffic is light or the system parameters fall farther away from the thresholds, e.g., when one type has a substantially larger cost, service rate, and proportion, then it is not necessary to use the G-$c\mu$ rule and in fact it could be better to use the optimal static policy, which does not require knowing the cost function precisely and is much simpler to implement.

7 Conclusions

In this paper, to answer some basic questions surrounding prioritization of certain customer groups in a service system, we studied a single-server queueing model with stationary Poisson arrivals of two types of customers with possibly distinct service time distributions and nonlinear waiting cost functions. When queue-waiting costs are nonlinear functions of time, it is known that the priority policy that minimizes the long-run average waiting costs would be state dependent, i.e., dependent on the durations of time customers in the queue have already spent waiting, in addition to their types. However, in practice, the most commonly employed queueing disciplines are still first-come-first-serve (FCFS) and type-based priority policies that give exclusive priority to one of the types of customers. In this paper, we compared these static policies in terms of their long-run average performance and derived several interesting insights. In particular, using the probabilistic analog of the mean value theorem, we obtained a complete ordering of the three policies (namely, FCFS, PF_1 that prioritizes type 1 customers, and PF_2 that prioritizes type 2 customers) for general cost functions under some mild existence conditions. To demonstrate how this result can be used in practice and to generate useful managerial insights, we then took a closer look at the case with polynomial cost functions, particularly the case with quadratic 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 will affect the characterization of optimal policies, and there will always be a type-based priority policy that performs at least as well as FCFS. However, we found that this is no longer the case when cost functions are quadratic and FCFS might perform better than prioritizing either one of the two types. In particular, we showed that the characterization of the best policy among FCFS, PF_1, and PF_2 depends on the rate of arrivals, proportion of each customer type in the population, first three moments of the service times, and cost parameters. For example, we found that if the two types of customers are similar in terms of the first two moments of their service times [mean service times], then the parameter region where FCFS is better than the two type-based priority policies enlarges with an increase in arrival rate [with higher service-time variability]. Hence, haphazardly replacing FCFS discipline with a type-based 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. One situation that we identified where it would be safe to replace FCFS discipline with a type-based priority policy is when the derivative of the cost function of one type is bounded from above (as in a linear cost function) and the other type has a quadratic cost function. In such a case, it is better to prioritize the type with quadratic cost under heavy traffic no matter how the service time distributions for the two types compare.

As a byproduct of our study on quadratic cost functions, we were also able to obtain some useful results on the problem of minimizing the variance of steady-state waiting times, which is widely accepted as a suitable performance measure to judge fairness of different queueing disciplines. In particular, for the case where the two types have equal means but possibly different higher moments, we showed that if the traffic intensity is above $\sqrt{2}/(1+\sqrt{2}) \approx 0.586$, then FCFS minimizes the variance of steady-state waiting times within the set of all static policies when neither type is more dominant in numbers. However, when the traffic intensity is below this threshold, then it is best to prioritize the type with smaller service-time variance.

We also conducted a numerical study to compare the performance of the best static policy identified through our analytical results with a benchmark state-dependent policy, namely, the G-$c\mu$ rule for M/M/1 queues with two types of customers under quadratic waiting costs. This study suggests that in most scenarios considered, using the best static policy would not result in significant differences in long-run average costs when compared with the G-$c\mu$ rule. More specifically, G-$c\mu$ performs better than the best static policy for a busy system when it is not clear which type is more "important" with respect to dominance in rates of cost and service. 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 it is not necessary to use the G-$c\mu$ rule since the best static policy, which is much easier to implement and which does not require precise knowledge on the waiting cost function, performs similarly or even slightly better.

References

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

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

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

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

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

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

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

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

D. R. Cox and W. L. Smith. *Queues.* Methuen, 1961.

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

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

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

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

D. Gross, J. F. Shortle, J. M. Thompson, and C. Harris. *Fundamentals of Queueing Theory.* John Wiley & Sons, Fourth edition, 2008.

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

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

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

N. K. Jaiswal. *Priority Queues.* Academic Press, 1968.

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

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

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

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

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

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

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

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

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

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

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

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

W. J. Sacco, D. M. Navin, K. E. Fiedler, I. Waddell, K. Robert, W. B. Long, and R. F. Buckman. Precise formulation and evidence-based application of resource-constrained triage. *Academic Emergency Medicine*, 12(8):759–770, 2005.

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

Z. Sun, N. T. Argon, and S. Ziya. Patient triage and prioritization under austere conditions. *Management Science, To Appear*, 2017.

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

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

R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Pearson College Division, 1989.

Appendix

In this Appendix, we provide proofs of results and other supplemental material that could not be presented in the main text due to space considerations.

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

$$\begin{aligned} C_\pi &= \sum_{i=1}^{2} \lim_{t\to\infty} \left(\frac{\sum_{k=1}^{n_i(t)} C_i(V_{i,k}^{\pi,x_0})}{n_i(t)} \right) \left(\frac{n_i(t)}{t} \right) \\ &= \sum_{i=1}^{2} \lim_{t\to\infty} \frac{\sum_{k=1}^{n_i(t)} C_i(V_{i,k}^{\pi,x_0})}{n_i(t)} \lim_{t\to\infty} \frac{n_i(t)}{t} = \sum_{i=1}^{2} \lambda p_i \lim_{n\to\infty} \frac{\sum_{k=1}^{n} C_i(V_{i,k}^{\pi,x_0})}{n}, \qquad (15) \end{aligned}$$

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

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} C_i(V_{i,k}^{\pi,x_0})}{n} = E\left[C_i(W_i^\pi)\right], \qquad (16)$$

which shows that (15) (and hence (1)) is equivalent to (2).

In the remainder of this proof, we drop the superscripts π and x_0 for notational convenience, and let T_{ik}, S_{ik} and D_{ik} be the arrival time, service time and departure time of the kth type i customer, respectively, under policy π and initial state x_0. Then, $V_{ik} = D_{ik} - T_{ik} - S_{ik}$ is the queue-waiting time for this customer. Note that $\{V_{ik}, k = 1, 2, , \ldots\}$ for each $i \in \{1, 2\}$ is a delayed regenerative process with nth regeneration happening at $N_{i,n}$ for $n = 0, 1, 2, \ldots$, where $N_{i,0} = 1$, and

$$N_{i,n} = \min\{k : k > N_{i,n-1}, V_{ik} = 0\}.$$

Note also that for each $i \in \{1, 2\}$, $\{C_i\left(V_{ik}\right), k = 1, 2, \ldots\}$ is a regenerative process with the same regeneration epoches as $\{V_{ik}, k = 1, 2, \ldots\}$. Then, by Theorem 13 of Chapter 2 and last paragraph of page 93 in Wolff (1989), (16) holds if $\sum_{k=1}^{N_{i,1}-1} |C_i(V_{ik})| < \infty$ with probability one, $E\left[N_{i,2} - N_{i,1}\right] < \infty$, and $E\left[\sum_{k=N_{i,1}}^{N_{i,2}-1} |C_i(V_{ik})|\right] < \infty$. We next complete the proof by showing that these three conditions hold.

When $\rho < 1$, the system is stable, i.e., it will return to the empty state within finite time with probability one and also the expected time it takes to return to the empty state is finite (see, e.g., Theorem 7.11 in Kulkarni (2009)). This implies that $N_{i,1} < \infty$ with probability one, $N_{i,2} - N_{i,1} <$

∞ with probability one, $V_{i,k} < \infty$ for any i and k with probability one and $E\left[N_{i,2} - N_{i,1}\right] < \infty$. At last, by Theorem B.5 (i) in El-Taha and Stidham Jr (1999), $E\left[\sum_{k=N_{i,1}}^{N_{i,2}-1} |C_i(V_{ik})|\right] = E\left[|C_i(W_i)|\right] E\left[N_{i,2} - N_{i,1}\right]$ is finite under the assumption that $E\left[|C_i(W_i)|\right]$ is finite. □

Proof of Lemma 3: We use sample path arguments to prove the stochastic inequalities. Let i be fixed to be either 1 or 2. Here type i and $3-i$ customers will be called priority and non-priority customers, respectively.

We index the customers by their arrival order to the system, and let s_j be the arriving time of customer j. Then, for customers l and j, where $j > l \geq 1$, we have $s_j > s_l$. Let t_j^π be the service starting time of customer j under policy π, then $t_j^\pi \geq s_j$. Let also V_j^π denote the waiting time of customer j under policy π, then $V_j^\pi = t_j^\pi - s_j$ for $j = 1, 2, \ldots$.

Under FCFS, we have $t_1^F < t_2^F < \cdots$ with probability one. Let j be the index of the first non-priority customer whose service starts when there are priority customers waiting, and k be the index of the first priority customer in the queue when j starts service under FCFS. Then, the customers indexed from j to $k-1$ are all non-priority customers. Note that $s_j < \cdots < s_{k-1} < s_k < t_j^F < \cdots < t_{k-1}^F < t_k^F$.

Consider a policy π that follows FCFS except that it serves customer k first, and then serves the non-priority customers $j, \ldots, k-1$. For the kth customer, who is a priority customer, $t_k^\pi = t_j^F < t_k^F$ and $V_k^\pi = t_k^\pi - s_k < t_k^F - s_k = V_k^F$. For $l = j, \ldots, k-1$, who are all non-priority customers, $t_l^\pi > t_l^F$ and $V_l^\pi = t_l^\pi - s_l > t_l^F - s_l = V_l^F$. For any $l \notin \{j, \ldots, k\}$, we have $V_l^\pi = V_l^F$.

If we keep changing the service order like this when there are non-priority customers starting service while priority customers are waiting in the queue, then we will eventually reach policy PF_i. This coupling argument then will yield $V_{i,n}^{PF_i} \leq_{st} V_{i,n}^F$ and $V_{3-i,n}^{PF_i} \geq_{st} V_{3-i,n}^F$ for $n \geq 1$. Since W_i^π is the steady-state waiting time for type i customers under policy π, then, as $n \to \infty$, $V_{i,n}^\pi \xrightarrow{d} W_i^\pi$ and $V_{3-i,n}^\pi \xrightarrow{d} W_{3-i}^\pi$, and hence, according to Theorem 1.A.3(d) in Shaked and Shanthikumar (2007), we have $W_i^{PF_i} \leq_{st} W^F$ and $W_{3-i}^{PF_i} \geq_{st} W^F$.

□

Proof of Theorem 1: We prove this result by comparing the costs directly.

(a) For $i \in \{1, 2\}$, Equation (2) yields $C_F \leq C_{PF_i}$ if and only if

$$\begin{aligned} & p_i\Big(E\big[C_i(W^F)\big] - E\big[C_i(W_i^{PF_i})\big]\Big) \leq p_{3-i}\Big(E\big[C_{3-i}(W_{3-i}^{PF_i})\big] - E\big[C_{3-i}(W^F)\big]\Big) \\ \Leftrightarrow \quad & p_i E\big[C_i'(U_i^{PF_i})\big]\big(E[W^F] - E[W_i^{PF_i}]\big) \leq p_{3-i} E\big[C_{3-i}'(U_{3-i}^{PF_i})\big]\big(E[W_{3-i}^{PF_i}] - E[W^F]\big) \end{aligned}$$

based on Lemma 1. Since

$$\frac{p_{3-i}\big(E[W_{3-i}^{PF_i}]-E[W^F]\big)}{p_i\big(E[W^F]-E[W_i^{PF_i}]\big)}=\frac{p_{3-i}\left(\frac{1}{(1-\rho)(1-\rho_i)}-\frac{1}{1-\rho}\right)}{p_i\left(\frac{1}{1-\rho}-\frac{1}{1-\rho_i}\right)}=\frac{p_{3-i}\rho_i}{p_i\rho_{3-i}}=\frac{\tau_i}{\tau_{3-i}},$$

we have $C_F \leq C_{PF_i}$ if and only if $a_i \leq b_i$.

(b) Equation (2) yields $C_{PF_1} \leq C_{PF_2}$ if and only

$$p_2\Big(E\big[C_2(W_2^{PF_1})\big]-E\big[C_2(W_2^{PF_2})\big]\Big)\leq p_1\Big(E\big[C_1(W_1^{PF_2})\big]-E\big[C_1(W_1^{PF_1})\big]\Big). \tag{17}$$

We have,

$$\begin{aligned}
&p_2\left(E\big[C_2(W_2^{PF_1})\big]-E\big[C_2(W_2^{PF_2})\big]\right)\\
=&p_2\left(E\big[C_2(W_2^{PF_1})\big]-E\big[C_2(W^F)\big]+E\big[C_2(W^F)\big]-E\big[C_2(W_2^{PF_2})\big]\right)\\
=&p_2\left(\big(E[W_2^{PF_1}]-E[W^F]\big)E\big[C_2'(U_2^{PF_1})\big]+\big(E[W_2^F]-E[W_2^{PF_2}]\big)E\big[C_2'(U_2^{PF_2})\big]\right) \qquad \text{(by (4))}\\
=&p_2\left(\frac{\lambda\rho_1\bar{\xi}}{2(1-\rho_1)(1-\rho)}E\big[C_2'(U_2^{PF_1})\big]+\frac{\lambda\rho_1\bar{\xi}}{2(1-\rho_2)(1-\rho)}E\big[C_2'(U_2^{PF_2})\big]\right)\\
=&\frac{\tau_2 p_2\lambda\rho_1\bar{\xi}}{2(1-\rho_1)(1-\rho_2)(1-\rho)}\left[(1-\rho_2)\frac{E\big[C_2'(U_2^{PF_1})\big]}{\tau_2}+(1-\rho_1)\frac{E\big[C_2'(U_2^{PF_2})\big]}{\tau_2}\right],\\
=&\frac{\rho_1\rho_2\bar{\xi}}{2(1-\rho_1)(1-\rho_2)(1-\rho)}\big[(1-\rho_2)b_1+(1-\rho_1)a_2\big].
\end{aligned} \tag{18}$$

Interchanging indices 1 and 2 in (18) yields

$$p_1\left(E\big[C_1(W_1^{PF_2})\big]-E\big[C_1(W_1^{PF_1})\big]\right)=\frac{\rho_1\rho_2\bar{\xi}}{2(1-\rho_1)(1-\rho_2)(1-\rho)}\big[(1-\rho_1)b_2+(1-\rho_2)a_1\big]. \tag{19}$$

Then, from (17), (18) and (19), we have $C_{PF_1} \leq C_{PF_2}$ if and only if $(1-\rho_1)(a_2-b_2) \leq (1-\rho_2)(a_1-b_1)$.

□

Proof of Corollary 2: (a) If $C_1'(t) \geq \tau_1 \max\{a_2, b_1\}$ for all $t \geq 0$, then for any non-negative random variable X, we have $E\left[C_1'(X)\right] \geq \tau_1 \max\{a_2, b_1\}$ when the expectation exists. Hence,

$$a_1=\frac{E\left[C_1'(U_1^{PF_1})\right]}{\tau_1}\geq\frac{\tau_1\max\{a_2,b_1\}}{\tau_1}\geq b_1,$$

which implies that $C_F \geq C_{PF_1}$ from Theorem 1(a). Similarly,

$$b_2 = \frac{E\left[C_1'(U_1^{PF_2})\right]}{\tau_1} \geq \frac{\tau_1 \max\{a_2, b_1\}}{\tau_1} \geq a_2,$$

which implies that $C_F \leq C_{PF_2}$ from Theorem 1(a).

(b) If $C_1'(t) \leq \tau_1 \min\{a_2, b_1\}$ for all $t \geq 0$, then we have $E\left[C_1'(X)\right] \leq \tau_1 \min\{a_2, b_1\}$ for any non-negative random variable X when the expectation exists. Then, $a_1 \leq b_1$ and $b_2 \leq a_2$, which implies that $C_F \leq C_{PF_1}$ and $C_F \geq C_{PF_2}$ from Theorem 1(a).

(c) If $\tau_1 a_2 \leq C_1'(t) \leq \tau_1 b_1$ for all $t \geq 0$, then we have $\tau_1 a_2 \leq E\left[C_1'(X)\right] \leq \tau_1 b_1$ for any non-negative random variable X when the expectation exists. Then, $a_i \leq b_i$ for $i \in \{1, 2\}$, which implies that $C_F \leq C_{PF_1}$ and $C_F \leq C_{PF_2}$ from Theorem 1(a).

□

Proof of Corollary 3: (a) Since $C_1'(t) \geq \max\{\alpha, \beta\} C_2'(t)$ for all $t \geq 0$, then for any non-negative random variable X, we have $E[C_1'(X)] \geq \max\{\alpha, \beta\} E[C_2'(X)]$ when the expectations exist. Consequently, for $X = U_1^{PF_1}$ we have

$$E[C_1'(U_1^{PF_1})] \geq \beta E[C_2'(U_1^{PF_1})] = \left(\frac{\tau_1}{\tau_2}\right) E[C_2'(U_2^{PF_1})] \Leftrightarrow a_1 \geq b_1,$$

and hence by Theorem 1(a), $C_{PF_1} \leq C_F$. Similarly, for $X = U_1^{PF_2}$ we have

$$E[C_1'(U_1^{PF_2})] \geq \alpha E[C_2'(U_1^{PF_2})] = \left(\frac{\tau_1}{\tau_2}\right) E[C_2'(U_2^{PF_2})] \Leftrightarrow b_2 \geq a_2,$$

and hence by Theorem 1(a), $C_F \leq C_{PF_2}$.

(b) Similar to part (a), since $C_1'(t) \leq \min\{\alpha, \beta\} C_2'(t)$ for all $t \geq 0$, we have $a_1 \leq b_1$ and $b_2 \leq a_2$. Thus, by Theorem 1(a), we have $C_{PF_2} \leq C_F \leq C_{PF_1}$.

(c) Similar to part (a), since $\alpha C_2'(t) \leq C_1'(t) \leq \beta C_2'(t)$ for all $t \geq 0$, we have $a_1 \leq b_1$ and $a_2 \leq b_2$, which implies that $C_F \leq C_{PF_1}$ and $C_F \leq C_{PF_2}$ by Theorem 1(a).

(d) Let W^π denote the steady-state waiting time for a randomly picked customer under policy

$\pi \in \{F, PF_1, PF_2\}$. By conditioning on the customer type, we have

$$E\left[C_2(W^\pi)\right] = p_1 E\left[C_2(W_1^\pi)\right] + p_2 E\left[C_2(W_2^\pi)\right]. \tag{20}$$

According to Theorem 2 in Vasicek (1977), if $C_2(\cdot)$ is convex, then we have

$$E[C_2(W^F)] \leq E[C_2(W^{PF_i})] \text{ for } i \in \{1,2\} \tag{21}$$

because service times are i.i.d. for all customers and they are subject to the same waiting cost function. From (20) and (21), we have

$$\begin{aligned} & p_i E\left[C_2(W^F)\right] + p_{3-i} E\left[C_2(W^F)\right] \leq p_i E\left[C_2(W_i^{PF_i})\right] + p_{3-i} E\left[C_2(W_{3-i}^{PF_i})\right] \\ \Leftrightarrow & \frac{E\left[C_2(W^F)\right] - E\left[C_2(W_i^{PF_i})\right]}{p_{3-i}} \leq \frac{E\left[C_2(W_{3-i}^{PF_i})\right] - E\left[C_2(W^F)\right]}{p_i} \text{ for } i \in \{1,2\}. \end{aligned} \tag{22}$$

Since $C_2(\cdot)$ is non-decreasing and $W_i^{PF_i} \leq_{st} W^F \leq_{st} W_{3-i}^{PF_i}$ from Lemma 3, we have

$$E\left[C_2(W^F)\right] - E\left[C_2(W_i^{PF_i})\right] \geq 0, \; E\left[C_2(W_{3-i}^{PF_i})\right] - E\left[C_2(W^F)\right] \geq 0, \text{ for } i \in \{1,2\}.$$

Then, (22) leads to

$$\left(\frac{p_i}{p_{3-i}}\right)\left(\frac{E\left[C_2(W^F)\right] - E\left[C_2(W_i^{PF_i})\right]}{E\left[C_2(W_{3-i}^{PF_i})\right] - E\left[C_2(W^F)\right]}\right) \leq 1 \text{ for } i \in \{1,2\}. \tag{23}$$

(Note that the numerator and denominator of (23) are positive for $i = 1$ and $i = 2$, respectively, by the assumption that $E\left[C_2'(U_1^{PF_1})\right] \neq 0$ and $E\left[C_2'(U_1^{PF_2})\right] \neq 0$, and Lemma 1.) Hence, by Lemmas 1 and 2, we have

$$\alpha = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{E[C_2'(U_2^{PF_2})]}{E[C_2'(U_1^{PF_2})]}\right) = \left(\frac{p_2}{p_1}\right)\left(\frac{E[C_2(W^F)] - E[C_2(W_2^{PF_2})]}{E[C_2(W_1^{PF_2})] - E[C_2(W^F)]}\right) \leq 1$$

and

$$\beta = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{E[C_2'(U_2^{PF_1})]}{E[C_2'(U_1^{PF_1})]}\right) = \left(\frac{p_2}{p_1}\right)\left(\frac{E[C_2(W_2^{PF_1})] - E[C_2(W^F)]}{E[C_2(W^F)] - E[C_2(W_1^{PF_1})]}\right) \geq 1.$$

□

We need new notation and Lemma 4 to prove Proposition 2. Let T_B denote the length of a busy period, which is defined as the length of time between the arrival of a customer at the empty system and the first subsequent time at which the system is again empty. We also define T_{B_j} as the length a busy period during which only type j customers arrive and are served for $j \in \{1,2\}$. Let $B(s)$ and $B_j(s)$ denote the respective LSTs of T_B and T_{B_j} for $s \geq 0$. For $i \in \{1,2\}$, we also let $\tilde{S}_i(s)$ denote the LST of the service time distribution for type i customers, and define $\tilde{S}(s) \equiv p_1\tilde{S}_1(s) + p_2\tilde{S}_2(s)$ for $s \geq 0$. Finally, let $\widetilde{W}^F(s)$ and $\widetilde{W}_i^{PF_j}(s)$ denote the respective LSTs of W^F and $W_i^{PF_j}$ for $i,j \in \{1,2\}$.

Lemma 4. *(Miller, 1960) For fixed $j \in \{1,2\}$, we have*

$$\widetilde{W}_j^{PF_j}(s) = \frac{(1-\rho)s + \lambda p_{3-j}\big(1-\tilde{S}_{3-j}(s)\big)}{s - \lambda p_j\Big(1-\tilde{S}_j(s)\Big)},\ \widetilde{W}^F(s) = \frac{(1-\rho)s}{s-\lambda\Big(1-\tilde{S}(s)\Big)},$$

and

$$\widetilde{W}_{3-j}^{PF_j}(s) = \widetilde{W}^F\big(\lambda p_j(1-B_j(s))+s\big) = \frac{1-\rho}{1-\frac{\lambda\left(1-\tilde{S}\left(\lambda p_j(1-B_j(s))+s\right)\right)}{\lambda p_j(1-B_j(s))+s}},$$

where $B_j(s)$ is the unique solution to $B_j(s) = \tilde{S}_j\big(s+\lambda p_j(1-B_j(s))\big)$ for $s > 0$ and $\lim_{s\to\infty} B_j(s) = 0$.

Proof of Proposition 2. Assumption 1 holds for type i for which $C_i(t)$ is in the form of (5) under policies FCFS, PF_1 and PF_2 if $E\left[\left(W^F\right)^l\right]$ and $E\left[\left(W_i^{PF_m}\right)^l\right]$ are finite for all $m \in \{1,2\}$ and $l = 1,2,\ldots,j(i)$. Note that by Lemma 3 and Theorem 1.A.3(a) of Shaked and Shanthikumar (2007), for $1 \leq l < \infty$, we have

$$E\left[\left(W_i^{PF_i}\right)^l\right] \leq E\left[\left(W^F\right)^l\right] \leq E\left[\left(W_i^{PF_{3-i}}\right)^l\right],$$

and thus we only need to prove that $E\left[\left(W_i^{PF_{3-i}}\right)^l\right]$ is finite. Note also that

$$E\left[\left(W_i^{PF_{3-i}}\right)^l\right] = (-1)^l \frac{d^l\widetilde{W}_i^{PF_{3-i}}(s)}{ds^l}\Big|_{s=0}, \tag{24}$$

and from Lemma 4, we have

$$\widetilde{W}_i^{PF_{3-i}}(s) = \widetilde{W}^F\big(\lambda p_{3-i}(1-B_{3-i}(s))+s\big).$$

Then, using Faa di Bruno's formula (see, e.g., Theorem 2 of Roman (1980)), (24) is finite if

$\frac{d^n\widetilde{W}^F(s)}{ds^n}|_{s=0}$ and $\frac{d^nB_{3-i}(s)}{ds^n}|_{s=0}$ are finite for all $n \leq l$, i.e., if the nth moment of W^F and $T_{B_{3-i}}$ are finite.

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

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

where $E\left[S^{j+1}\right]$ is the $(j+1)$st moment of service time of a randomly picked customer with a service time LST of $\tilde{S}(s)$. Hence, $E\left[\left(W^F\right)^n\right]$ is finite if $\rho < 1$ and the first $n+1$ moments of service times of both types are finite. Besides, from Theorem 1 of Ghahramani and Wolff (1989), the nth moment of the busy period for a single-class queue is finite if and only if the nth moment of the service times is finite. Thus, $E\left[\left(W_i^{PF_{3-i}}\right)^l\right]$ is finite if $\rho < 1$ and the first $l+1$ moments of service times are finite.

□

Proof of Proposition 3. We first drive expressions (8) and (9) using definitions of a_i and b_i given in Theorem 1. For some $i, k, m \in \{1, 2\}$, we have

$$E\left[C_i'(U_k^{PF_m})\right] = 2k_iE\left[U_k^{PF_m}\right] + h_i = k_i\left(\frac{E\left[(W^F)^2\right] - E\left[(W_k^{PF_m})^2\right]}{E[W^F] - E[W_k^{PF_m}]}\right) + h_i,$$

from Equations (6) and (7). The expected waiting times have been given in Lemma 2, and the second moments can be obtained from Gross et al. (2008) and Miller (1960):

$$E\left[(W^F)^2\right] = \frac{\lambda\bar{\zeta}}{3(1-\rho)} + \frac{\lambda^2\bar{\xi}^2}{2(1-\rho)^2},\ E\left[(W_k^{PF_k})^2\right] = \frac{\lambda\bar{\zeta}}{3(1-\rho_k)} + \frac{\lambda^2p_k\xi_k\bar{\xi}}{2(1-\rho_k)^2},$$

and

$$E\left[(W_{3-k}^{PF_k})^2\right] = \frac{\lambda\bar{\zeta}}{3(1-\rho_k)^2(1-\rho)} + \frac{\lambda^2\bar{\xi}^2}{2(1-\rho_k)^2(1-\rho)^2} + \frac{\lambda^2p_k\xi_k\bar{\xi}}{2(1-\rho_k)^3(1-\rho)}.$$

Then, we find

$$\begin{aligned}E[W^F] - E[W_k^{PF_k}] &= \frac{\lambda^2p_{3-k}\tau_{3-k}\bar{\xi}}{2(1-\rho_k)(1-\rho)},\\ E[W_{3-k}^{PF_k}] - E[W^F] &= \frac{\lambda^2p_k\tau_k\bar{\xi}}{2(1-\rho_k)(1-\rho)},\end{aligned}$$

$$
\begin{aligned}
&E\left[\left(W^F\right)^2\right] - E\left[\left(W_k^{PF_k}\right)^2\right] \\
&= \frac{\rho_{3-k}\lambda\bar{\zeta}}{3(1-\rho_k)(1-\rho)} + \frac{\lambda^2\bar{\xi}}{2(1-\rho_k)(1-\rho)}\left[\frac{(1-\rho_k)(p_1\xi_1+p_2\xi_2)}{1-\rho} - \frac{p_k\xi_k(1-\rho)}{1-\rho_k}\right] \\
&= \frac{\lambda^2 p_{3-k}\tau_{3-k}\bar{\zeta}}{3(1-\rho_k)(1-\rho)} + \frac{\lambda^2 p_{3-k}\bar{\xi}}{2(1-\rho_k)(1-\rho)}\left[\frac{p_k\xi_k\lambda\tau_{3-k}(2-\rho-\rho_k)}{(1-\rho)(1-\rho_k)} + \frac{\xi_{3-k}(1-\rho_k)}{1-\rho}\right],
\end{aligned}
$$

$$
E\left[\left(W_{3-k}^{PF_k}\right)^2\right] - E\left[\left(W^F\right)^2\right] = \left[\frac{\bar{\zeta}}{3} + \frac{\lambda\bar{\xi}^2}{2(1-\rho)}\right]\frac{\lambda\rho_k(2-\rho_k)}{(1-\rho)(1-\rho_k)^2} + \frac{\lambda^2 p_k\xi_k\bar{\xi}}{2(1-\rho_k)^3(1-\rho)}
$$
$$
= \left[\frac{2\bar{\zeta}(2-\rho_k)}{3\bar{\xi}(1-\rho_k)} + \frac{\lambda\bar{\xi}(2-\rho_k)}{(1-\rho)(1-\rho_k)}\right]\frac{\lambda\rho_k\bar{\xi}}{2(1-\rho_k)(1-\rho)} + \frac{\lambda\rho_k\xi_k\bar{\xi}}{2\tau_k(1-\rho_k)^3(1-\rho)}.
$$

Hence,

$$
\begin{aligned}
E\big[C_i'(U_k^{PF_k})\big] &= k_i\left[\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{p_k\xi_k\lambda(2-\rho-\rho_k)}{(1-\rho)(1-\rho_k)} + \frac{\xi_{3-k}(1-\rho_k)}{\tau_{3-k}(1-\rho)}\right] + h_i \\
&= k_i\left[\frac{2\bar{\zeta}}{3\bar{\xi}} + \lambda p_k\xi_k\left(\frac{1}{1-\rho} + \frac{1}{1-\rho_k}\right) + \lambda p_{3-k}\xi_{3-k}\left(\frac{1}{1-\rho} + \frac{1}{\rho_{3-k}}\right)\right] + h_i \\
&= k_i\left[\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_k\xi_k}{1-\rho_k} + \frac{\xi_{3-k}}{\tau_{3-k}}\right] + h_i, \qquad (25)
\end{aligned}
$$

and

$$
E\big[C_i'(U_{3-k}^{PF_k})\big] = k_i\left[\frac{2\bar{\zeta}}{3\bar{\xi}}\left(1+\frac{1}{1-\rho_k}\right) + \frac{\lambda\bar{\xi}}{1-\rho}\left(1+\frac{1}{1-\rho_k}\right) + \frac{\xi_k}{\tau_k(1-\rho_k)^2}\right] + h_i, \qquad (26)
$$

which yield Equations (8) and (9).

We next show that $a_i < b_{3-i}$ for $i = 1, 2$. For $i \in \{1, 2\}$, Equations (8) and (9) yield

$$
\begin{aligned}
b_{3-i} - a_i = \left(\frac{k_i}{\tau_i}\right)\Bigg[&\frac{2\bar{\zeta}}{3\bar{\xi}(1-\rho_{3-i})} + \frac{\lambda\bar{\xi}\rho_{3-i}}{1-\rho}\left(\frac{1}{1-\rho_{3-i}} + \frac{1}{1-\rho_i}\right) \\
&+\lambda p_{3-i}\xi_{3-i}\left(\frac{1}{(1-\rho_{3-i})^2} + \frac{1}{1-\rho_i} + \frac{1}{1-\rho_{3-i}}\right)\Bigg], \qquad (27)
\end{aligned}
$$

which is positive because $\rho, \rho_1, \rho_2 < 1$ and all moments of service times are positive. Therefore, when both cost functions are quadratic, if $a_i > b_i$ for some $i \in \{1, 2\}$ (and thus $a_{3-i} \le b_i < a_i \le b_{3-i}$), then PF_i is the best among FCFS, PF_1 and PF_2 according to Corollary 1(b); otherwise ($a_1 \le b_1$ and $a_2 \le b_2$), FCFS is the best among FCFS, PF_1, and PF_2 by Corollary 1(a). □

Proof of Corollary 4. The expressions for A and B and the characterization of the optimal policy follow from (10). We next prove that $A < B$. Let $G_i(\lambda) = \frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\lambda\bar{\tau}} + \frac{\lambda p_i \xi_i}{1-\lambda p_i \tau_i}$ and $X_i = \frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\lambda\bar{\tau}} + \frac{\xi_i}{\tau_i}$ for $i \in \{1,2\}$. Then, we have

$$A = \frac{X_1 + \frac{\lambda p_2 \xi_2}{1-\rho_2}}{X_2 + \frac{G_2(\lambda)+\lambda p_2 \xi_2}{1-\rho_2}},\ B = \frac{X_1 + \frac{G_1(\lambda)+\lambda p_1 \xi_1}{1-\rho_1}}{X_2 + \frac{\lambda p_1 \xi_1}{1-\rho_1}}.$$

Note that for $i \in \{1,2\}$,

$$\frac{G_i(\lambda) + \lambda p_i \xi_i}{1-\rho_i} > \frac{G_i(\lambda)}{1-\rho_i} > \frac{\lambda\bar{\xi}}{(1-\rho_i)(1-\rho)} > \frac{\lambda p_{3-i}\xi_{3-i}}{1-\rho_{3-i}},$$

where the last inequality follows from the fact that $\bar{\xi} > p_{3-i}\xi_{3-i}$ and $(1-\rho_i)(1-\rho) < 1-\rho < 1-\rho_{3-i}$. Hence, we have

$$A < \frac{X_1 + \frac{\lambda p_2 \xi_2}{1-\rho_2}}{X_2 + \frac{\lambda p_1 \xi_1}{1-\rho_1}} < B.$$

□

Proof of Proposition 4. (a) From the expression of A in the proof of Corollary 4, we have,

$$\begin{aligned}\frac{\partial A}{\partial \lambda} &= \frac{G_2'(\lambda)\left(\frac{2-\rho_2}{1-\rho_2}G_2(\lambda) + \frac{\xi_2}{\tau_2}\right) - \left(G_2(\lambda) + \frac{\xi_1}{\tau_1}\right)\left(\frac{p_2\tau_2}{(1-\rho_2)^2}G_2(\lambda) + \frac{2-\rho_2}{1-\rho_2}G_2'(\lambda)\right)}{\left(\frac{2-\rho_2}{1-\rho_2}G_2(\lambda) + \frac{\xi_2}{\tau_2}\right)^2} \\ &= \frac{G_2'(\lambda)\left(\frac{\xi_2}{\tau_2} - \frac{(2-\rho_2)\xi_1}{(1-\rho_2)\tau_1}\right) - \left(G_2(\lambda) + \frac{\xi_1}{\tau_1}\right)\frac{p_2\tau_2}{(1-\rho_2)^2}G_2(\lambda)}{\left(\frac{2-\rho_2}{1-\rho_2}G_2(\lambda) + \frac{\xi_2}{\tau_2}\right)^2} < 0\end{aligned}$$

if and only if

$$G_2'(\lambda)\left(\frac{\xi_2}{\tau_2} - \frac{(2-\rho_2)\xi_1}{(1-\rho_2)\tau_1}\right) - \left(G_2(\lambda) + \frac{\xi_1}{\tau_1}\right)\frac{p_2\tau_2}{(1-\rho_2)^2}G_2(\lambda) < 0. \tag{28}$$

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

$$G_i'(\lambda) = \frac{\bar{\xi}}{(1-\lambda\bar{\tau})^2} + \frac{p_i\xi_i}{(1-\lambda p_i \tau_i)^2} > 0.$$

Then, (28) is equivalent to

$$\frac{\xi_2}{\tau_2} - \frac{(2-\rho_2)\xi_1}{(1-\rho_2)\tau_1} < \frac{\frac{p_2\tau_2}{(1-\rho_2)^2}\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_2\xi_2}{1-\rho_2}\right)\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_2\xi_2}{1-\rho_2} + \frac{\xi_1}{\tau_1}\right)}{\frac{\bar{\xi}}{(1-\rho)^2} + \frac{p_2\xi_2}{(1-\rho_2)^2}}.$$

(b) Similarly,

$$\begin{aligned}\frac{\partial B}{\partial \lambda} &= \frac{\left(\frac{p_1\tau_1}{(1-\rho_1)^2}G_1(\lambda) + \frac{2-\rho_1}{1-\rho_1}G_1'(\lambda)\right)\left(G_1(\lambda) + \frac{\xi_2}{\tau_2}\right) - G_1'(\lambda)\left(\frac{2-\rho_1}{1-\rho_1}G_1(\lambda) + \frac{\xi_1}{\tau_1}\right)}{\left(G_1(\lambda) + \frac{\xi_2}{\tau_2}\right)^2} \\ &= \frac{\frac{p_1\tau_1}{(1-\rho_1)^2}G_1(\lambda)\left(G_1(\lambda) + \frac{\xi_2}{\tau_2}\right) + G_1'(\lambda)\left(\frac{(2-\rho_1)\xi_2}{(1-\rho_1)\tau_2} - \frac{\xi_1}{\tau_1}\right)}{\left(G_1(\lambda) + \frac{\xi_2}{\tau_2}\right)^2} > 0\end{aligned}$$

if and only if

$$\frac{\xi_1}{\tau_1} - \frac{(2-\rho_1)\xi_2}{(1-\rho_1)\tau_2} < \frac{\frac{p_1\tau_1}{(1-\rho_1)^2}\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_1\xi_1}{1-\rho_1}\right)\left(\frac{2\bar{\zeta}}{3\bar{\xi}} + \frac{\lambda\bar{\xi}}{1-\rho} + \frac{\lambda p_1\xi_1}{1-\rho_1} + \frac{\xi_2}{\tau_2}\right)}{\frac{\bar{\xi}}{(1-\rho)^2} + \frac{p_1\xi_1}{(1-\rho_1)^2}}.$$

(c) Follows directly from the application of L'Hopital's rule. □

Proof of Proposition 5. (a) When service times are i.i.d., for notational convenience we drop the subscript from all parameters related to the service time distribution, i.e., τ_i, ξ_i and ζ_i. Then,

$$A = \frac{\frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho} + \frac{\xi}{\tau(1-\rho_2)}}{\frac{2-\rho_2}{1-\rho_2}\left(\frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho}\right) + \frac{\xi}{\tau(1-\rho_2)^2}}, \; B = \frac{\frac{2-\rho_1}{1-\rho_1}\left(\frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho}\right) + \frac{\xi}{\tau(1-\rho_1)^2}}{\frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho} + \frac{\xi}{\tau(1-\rho_1)}}.$$

Let $M \equiv \frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho}$, which is positive and not changing with respect to p_i for $i \in \{1, 2\}$, then we have

$$\frac{\partial A}{\partial p_2} = \frac{-\tau M^2 - \frac{\rho_2\xi}{(1-\rho_2)}M}{\frac{(1-\rho_2)^2}{\lambda}\left[\frac{2-\rho_2}{1-\rho_2}M + \frac{\xi}{\tau(1-\rho_2)^2}\right]^2} < 0.$$

Similarly, computing the partial derivative of B with respect to p_1, we have

$$\frac{\partial B}{\partial p_1} = \frac{\tau M^2 + \frac{\rho_1\xi}{(1-\rho_1)}M}{\frac{(1-\rho_1)^2}{\lambda}\left[M + \frac{\xi}{\tau(1-\rho_1)}\right]^2} > 0.$$

(b) In heavy traffic, i.e., as $\lambda \to 1/\bar{\tau}$, the expressions of A and B are given in Proposition 4. By letting $p_1 \to 0$, we have $p_2 \to 1$, and hence, we have $\lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 0} A = 0$ and $\lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 0} B = 2$. Similarly, by letting $p_1 \to 1$, we have $p_2 \to 0$, and then $\lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 1} A = 1/2$ and $\lim_{\lambda \to 1/\bar{\tau},\ p_1 \to 0} B = \infty$.

□

Proof of Proposition 6. When service times are exponential, $\zeta_i = 6\tau_i^3$ and $\xi_i = 2\tau_i^2$ for $i \in \{1, 2\}$, and hence we have,

$$A_{exp} = \frac{\frac{p_1\tau_1^3+p_2\tau_2^3}{p_1\tau_1^2+p_2\tau_2^2} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_2\tau_2^2}{1-\rho_2} + \tau_1}{\frac{2-\rho_2}{1-\rho_2}\left(\frac{p_1\tau_1^3+p_2\tau_2^3}{p_1\tau_1^2+p_2\tau_2^2} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_2\tau_2^2}{1-\rho_2}\right) + \tau_2},$$

$$B_{exp} = \frac{\frac{2-\rho_1}{1-\rho_1}\left(\frac{p_1\tau_1^3+p_2\tau_2^3}{p_1\tau_1^2+p_2\tau_2^2} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_1\tau_1^2}{1-\rho_1}\right) + \tau_1}{\frac{p_1\tau_1^3+p_2\tau_2^3}{p_1\tau_1^2+p_2\tau_2^2} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_1\tau_1^2}{1-\rho_1} + \tau_2}.$$

When service times are deterministic, $\zeta_i = \tau_i^3$ and $\xi_i = \tau_i^2$ for $i \in \{1, 2\}$, and then we have

$$A_{det} = \frac{\frac{2\left(p_1\tau_1^3+p_2\tau_2^3\right)}{3\left(p_1\tau_1^2+p_2\tau_2^2\right)} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_2\tau_2^2}{1-\rho_2} + \tau_1}{\frac{2-\rho_2}{1-\rho_2}\left(\frac{2\left(p_1\tau_1^3+p_2\tau_2^3\right)}{3\left(p_1\tau_1^2+p_2\tau_2^2\right)} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_2\tau_2^2}{1-\rho_2}\right) + \tau_2},$$

$$B_{det} = \frac{\frac{2-\rho_1}{1-\rho_1}\left(\frac{2\left(p_1\tau_1^3+p_2\tau_2^3\right)}{3\left(p_1\tau_1^2+p_2\tau_2^2\right)} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_1\tau_1^2}{1-\rho_1}\right) + \tau_1}{\frac{2\left(p_1\tau_1^3+p_2\tau_2^3\right)}{3\left(p_1\tau_1^2+p_2\tau_2^2\right)} + \frac{\lambda\left(p_1\tau_1^2+p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_1\tau_1^2}{1-\rho_1} + \tau_2}.$$

(a) For notational simplicity, for $i \in \{1, 2\}$, we let

$$\begin{aligned} M_{exp}^{(i)} &\equiv \frac{p_1\tau_1^3 + p_2\tau_2^3}{p_1\tau_1^2 + p_2\tau_2^2} + \frac{\lambda\left(p_1\tau_1^2 + p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_i\tau_i^2}{1-\rho_i}, \\ M_{det}^{(i)} &\equiv \frac{2\left(p_1\tau_1^3 + p_2\tau_2^3\right)}{3\left(p_1\tau_1^2 + p_2\tau_2^2\right)} + \frac{\lambda\left(p_1\tau_1^2 + p_2\tau_2^2\right)}{1-\rho} + \frac{\lambda p_i\tau_i^2}{1-\rho_i}, \end{aligned}$$

where $M_{exp}^{(i)} > M_{det}^{(i)}$. Taking the difference of A_{exp} and A_{det}, we have

$$A_{exp} - A_{det} = \frac{\left(\tau_2 - \left(\frac{2-\rho_2}{1-\rho_2}\right)\tau_1\right)\left(M_{exp}^{(2)} - M_{det}^{(2)}\right)}{\left(\left(\frac{2-\rho_2}{1-\rho_2}\right)M_{exp}^{(2)} + \tau_2\right)\left(\left(\frac{2-\rho_2}{1-\rho_2}\right)M_{det}^{(2)} + \tau_2\right)}.$$

Hence, $A_{exp} \leq A_{det}$ if and only if $\tau_2 \leq \frac{2-\rho_2}{1-\rho_2}\tau_1$.

(b) Taking the difference of B_{exp} and B_{det}, we have

$$B_{exp} - B_{det} = \frac{\left(\left(\frac{2-\rho_1}{1-\rho_1}\right)\tau_2 - \tau_1\right)\left(M_{exp}^{(1)} - M_{det}^{(1)}\right)}{\left(M_{det}^{(1)} + \tau_2\right)\left(M_{exp}^{(1)} + \tau_2\right)}.$$

Hence, $B_{exp} \geq B_{det}$ if and only if $\tau_1 \leq \frac{2-\rho_1}{1-\rho_1}\tau_2$.

□

Proof of Proposition 7. The expressions for α and β can be obtained from Corollary 3 and Equations (25) and (26). Let $A_d[B_d]$ and $A_n[B_n]$ denote the denominator and numerator of $A[B]$, as given in Corollary 4, respectively. Then,

$$\alpha = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{k_2 A_n + h_2}{k_2 A_d + h_2}\right), \ \beta = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{k_2 B_n + h_2}{k_2 B_d + h_2}\right). \tag{29}$$

(a) We have $A < B$ and hence $A_n B_d - B_n A_d < 0$. Besides,

$$B_d + A_n - A_d - B_n = -\frac{G_1(\lambda)}{1-\rho_1} - \frac{G_2(\lambda)}{1-\rho_2} < 0.$$

Hence,

$$\alpha - \beta = \left(\frac{\tau_1}{\tau_2}\right)\left(\frac{k_2^2(A_n B_d - B_n A_d) + h_2 k_2 (B_d + A_n - A_d - B_n)}{(k_2 A_d + h_2)(k_2 B_d + h_2)}\right) < 0.$$

(b) From Corollary 4, if (11) holds, then

$$\frac{\partial A}{\partial \lambda} = \frac{A_n' A_d - A_d' A_n}{A_d^2} < 0,$$

where A_d' and A_n' denote the partial derivatives of A_d and A_n with respect to λ, respectively. Then, $A_n' A_d - A_d' A_n < 0$. Besides, the difference $A_d - A_n = \frac{G_2(\lambda)}{1-\rho_2} + \frac{\xi_2}{\tau_2} - \frac{\xi_1}{\tau_1}$ increases in λ since $G_2(\lambda)$ increases and $1-\rho_2$ decreases in λ (see the proof of Corollary 4). Then, $A_d' > A_n'$.

$$\frac{\partial \alpha}{\partial \lambda} = \left(\frac{k_2 \tau_1}{\tau_2}\right)\left(\frac{k_2(A_n' A_d - A_d' A_n) + h_2(A_n' - A_d')}{(k_2 A_d + h_2)^2}\right) < 0.$$

Thus, α decreases as λ increases.

Similarly, if (12) holds, then $B_n' B_d - B_d' B_n > 0$ from Corollary 4, where B_d' and B_n' denote the partial derivatives of B_d and B_n with respect to λ, respectively. Besides, the difference $B_n - B_d = \frac{G_1(\lambda)}{1-\rho_1} + \frac{\xi_1}{\tau_1} - \frac{\xi_2}{\tau_2}$ increases in λ since $G_1(\lambda)$ increases and $1-\rho_1$ decreases in λ (see the proof of Corollary 4). Then, $B_n' > B_d'$.

$$\frac{\partial \beta}{\partial \lambda} = \left(\frac{k_2\tau_1}{\tau_2}\right)\left(\frac{k_2(B_n' B_d - B_d' B_n) + h_2(B_n' - B_d')}{(k_2 B_d + h_2)^2}\right) > 0.$$

Thus, β increases as λ increases.

When $h_2 = 0$, we know from (29) that $\alpha = A\tau_1/\tau_2$ and $\beta = B\tau_1/\tau_2$. Then, Proposition 4 directly provides the necessity of conditions (11) and (12). The limits for α and β as $\lambda \to 1/\bar{\tau}$ follows from (29), application of L'Hospital's rule and part (c) of Proposition 4.

(c) Let $\bar{A}_n[\bar{B}_n]$ and $\bar{A}_d[\bar{B}_d]$ denote the numerator and denominator of $A[B]$, as given in the proof of Proposition 5, then

$$\alpha = \frac{k_2\bar{A}_n + h_2}{k_2\bar{A}_d + h_2},\ \beta = \frac{k_2\bar{B}_n + h_2}{k_2\bar{B}_d + h_2}.$$

From Proposition 5, A and B both increase in p_1, then $\bar{A}_n'\bar{A}_d - \bar{A}_d'\bar{A}_n > 0$ and $\bar{B}_n'\bar{B}_d - \bar{B}_d'\bar{B}_n > 0$, where A_n', A_d', B_n', B_d' denote the partial derivatives of each quantity with respect to p_1. Besides, the difference $\bar{A}_d - \bar{A}_n = \frac{1}{1-\rho_2}\left(\frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho} + \frac{\lambda p_2 \xi}{1-\rho_2}\right)$ increases in p_2, and hence decreases in p_1, which implies that $\bar{A}_d' < \bar{A}_n'$. Then, we have

$$\frac{\partial \alpha}{\partial p_1} = \frac{k_2^2(\bar{A}_n'\bar{A}_d - \bar{A}_d'\bar{A}_n) + k_2 h_2(\bar{A}_n' - \bar{A}_d')}{(k_2\bar{A}_d + h_2)^2} > 0.$$

Similarly, the difference $\bar{B}_n - \bar{B}_d = \frac{1}{1-\rho_1}\left(\frac{2\zeta}{3\xi} + \frac{\lambda\xi}{1-\rho} + \frac{\lambda p_1 \xi}{1-\rho_1}\right)$ increases in p_1, and hence $\bar{B}_n' > \bar{B}_d'$. Then, we have $\frac{\partial \beta}{\partial p_1} > 0$.

□

Example 3. Suppose that $C_2(t) = k_2 t^2 + h_2 t$ for $h_2, k_2 \geq 0$. We next apply Corollary 3 and Proposition 7 to three different waiting cost functions for type 1 customers.

(i) Let $C_1(t) = h_1 t$ for $t \geq 0$, where h_1 is positive. We compare $C_1'(t) = h_1$ with $\alpha C_2'(t)$ and $\beta C_2'(t)$ for all $t \geq 0$, where $C_2'(t) = 2k_2 t + h_2$. Since $C_1'(t)$ is fixed and $C_2'(t)$ is increasing without any bound, the only applicable case from Corollary 3 is that $C_1'(t) \leq \alpha C_2'(t)$ for all $t \geq 0$, which is true if and only if $h_1 \leq \alpha h_2$. Hence, by applying Corollary 3, we know that PF_2 is better than FCFS and PF_1 if $h_1 \leq \alpha h_2$. This means that Corollary 3 provides a partial comparison of the three policies. On the other hand, for this case, Example 1(i) showed that

Corollary 2 lead to a complete characterization. (Indeed, one can show that $\alpha h_2 < \tau_1 a_2$ for qudratic cost functions.) Hence, Corollary 2 is more useful for this example.

(ii) When $C_1(t) = h_1 \ln(t+1)$ for $t \geq 0$ and positive constant h_1, we have $C_1'(t) = h_1/(t+1)$. As $t \to \infty$, we have $C_1'(t) \to 0$ and $C_2'(t) \to \infty$. Hence, the only applicable case from Corollary 3 is that $C_1'(t) \leq \alpha C_2'(t)$ for all $t \geq 0$, which is true if and only if $h_1 \leq \alpha h_2$. By applying Corollary 3, we find that if $h_1 \leq \alpha h_2$, then PF_2 is the best. In this example, both Corollaries 2 and 3 provide upper bounds on h_1 when PF_2 is the best, and since we can show that $\tau_1 a_2 > h_2 \alpha$, the bound from Corollary 2 is better than that from Corollary 3.

(iii) When $C_1(t) = k_1(e^{h_1 t} - 1)$ for $t \geq 0$ and positive constants k_1 and h_1, we have $C_1'(t) = k_1 h_1 e^{h_1 t}$. Since $C_1'(t)$ is exponential and $C_2'(t)$ is linear, $C_1'(t)$ will be greater than $C_2'(t)$ for sufficiently large t, and thus the only applicable case from Corollary 3 is that $C_1'(t) \geq \beta C_2'(t)$ for all $t \geq 0$, which is true if and only if the following condition holds (the proof is provided below):

$$k_1 \geq \frac{\beta}{h_1} \max\left\{h_2, \frac{2k_2}{h_1}\right\} e^{\min\left\{\frac{h_2 h_1}{2k_2} - 1, 0\right\}}. \tag{30}$$

In this example, both Corollaries 2 and 3 provide conditions under which PF_1 is the best. Whether Corollary 2 or 3 is better depends on the system parameters. To be more specific, for this example, Corollary 2 is more useful (in that it provides a weaker condition on the optimality of PF_1) if $\tau_1 b_1 h_1 < 2k_2 \beta e^{\left(\frac{h_2 h_1}{2k_2} - 1\right)}$ and $h_2 < \frac{2k_2}{h_1}$, and Corollary 3 is more useful otherwise.

Proof of Example 3(iii). Let $f(t) = C_1'(t) - \beta C_2'(t) = k_1 h_1 e^{h_1 t} - \beta(2k_2 t + h_2)$ for $t \geq 0$. We need to find conditions so that $f(t) \geq 0$ for all $t \geq 0$. First we find that $f(t)$ is convex in t for $t \geq 0$ by the second derivative test, so there is a global minimum for $t \geq 0$. We solve for $f'(t) = 0$ and we have a stationary point $t^* = \frac{1}{h_1} \ln\left(\frac{2\beta k_2}{k_1 h_1^2}\right)$. We have the following two cases:

Case 1: If $t^* \leq 0$, i.e., $k_1 \geq \frac{2\beta k_2}{h_1^2}$, then the minimum happens at $t = 0$, and hence we need $f(0) = k_1 h_1 - h_2 \beta \geq 0$, which is equivalent to

$$k_1 \geq \frac{h_2 \beta}{h_1}. \tag{31}$$

Case 2: If $t^* > 0$, i.e., $k_1 < \frac{2\beta k_2}{h_1^2}$, the minimum happens at t^* and hence we need $f(t^*) =$

$\frac{2\beta k_2}{h_1} - \frac{2\beta k_2}{h_1} \ln\left(\frac{2\beta k_2}{k_1 h_1^2}\right) - \beta h_2 \geq 0$, which is equivalent to

$$k_1 \geq \frac{2\beta k_2}{h_1^2} e^{-\left(1 - \frac{h_2 h_1}{2k_2}\right)}. \tag{32}$$

Note that if $h_2 \geq \frac{2k_2}{h_1}$, then condition (32) implies that $k_1 \geq \frac{2\beta k_2}{h_1^2}$, which contradicts with the condition for Case 2. Hence, if $h_2 \geq \frac{2k_2}{h_1}$, then the condition must be (31). If $h_2 < \frac{2k_2}{h_1}$ and $k_1 \geq \frac{2\beta k_2}{h_1^2}$ (i.e., Case 1 is true), then condition (31) is automatically satisfied. Finally, if $h_2 < \frac{2k_2}{h_1}$ and $k_1 < \frac{2\beta k_2}{h_1^2}$ (i.e., Case 2 is true), then the condition is (32). Combining these last three sentences implies that $f(t) \geq 0$ if and only if (30) holds. □