RUTCOR
RESEARCH
REPORT

Class treatment in Queueing Systems: Discrimination and Fairness Aspects

David Raz [a] Benjamin Avi-Itzhak [b]
Hanoch Levy [c]

RRR 36-2005, November, 2005

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, davidraz@post.tau.ac.il
[b]RUTCOR, Rutgers, the State University of New Jersey, 640 Bartholomew Road, Piscataway, NJ 08854-8003, USA, aviitzha@rutcor.rutgers.edu
[c]School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, hanoch@cs.tau.ac.il

Rutcor Research Report
RRR 36-2005, November, 2005

Class treatment in Queueing Systems: Discrimination and Fairness Aspects

David Raz Benjamin Avi-Itzhak Hanoch Levy

Abstract. Customer classification and prioritization are commonly used in applications to provide queue preferential service. Their influence on queuing systems has been thoroughly studied from the delay distribution perspective. However, the fairness aspects, which are inherent to any preferential system and highly important to customers, have hardly been studied and not been quantified to date. We use RAQFM to analyze such systems and derive their relative fairness values. Based on RAQFM we further introduce a new metrics to evaluate the discrimination experienced by various classes. Very common practices for treating classes, in public facilities as well as in computer systems, are those of *class prioritization* and *dedication of resources to classes*. We analyze both practices and study their effect on the fairness of the whole system as well as on the discrimination they inflict on the various classes. Some of our results are: 1) If service order is independent of service times, short jobs are negatively discriminated, 2) Assigning higher priority to short jobs often increases the system fairness, but not always, 3) Dedicating equal amount of resources to classes with different service times negatively discriminates the longest service time class, at least in the case of the $GI/M/1$ queue. Practitioners can use the derived results as well as the measures studied to weigh efficiency aspects versus fairness aspects in controlling their queueing systems. Accounting for the discrimination factor is especially important when customer classification is based on personal properties, such as customer gender.

Acknowledgements: This work was supported in part by grant 380-801 from the Israeli Ministry of Science and Technology

1 Introduction

1.1 Overview

Customer classification and prioritization are common mechanisms used in a large variety of daily queueing situations. One of the major reasons for using priorities and preferential service is that of fairness, that is, the wish to make the system operation "fair". Fairness among customers/jobs is a crucial and fundamental issue for queueing systems. A recent Experimental Psychology study by Rafaeli et al. [19, 20], where attitude of people in queues was studied, shows that fairness in the queue is very important to people, perhaps not less than the wait itself. In fact, our own observation is that perhaps the major reason for using an ordered queue at all is the wish to provide fair service and fair waiting to the customers.

Despite this fact, the fairness aspects of prioritization and classification in queueing systems have not been studied quantitatively to date, and their effect cannot be accounted for in the queueing analysis of many daily life applications. A striking and well known example are public restrooms where it is common to observe drastic imbalance between the queues of the two genders. This issue, which has been addressed in some places by legislative bodies, has not been examined yet via queue fairness analysis.

The objective of this work is to use a quantitative model for measuring the relative fairness of priority and classification systems. Such measurements can be used to quantitatively account for fairness when considering alternative designs. This can enhance the existing design approaches in which efficiency (e.g., utilization and delays) is accounted for quantitatively, while fairness is accounted for only in a qualitative way. To carry out the analysis we use the *Resource Allocation Queueing Fairness Measure (RAQFM)* introduced recently (see Raz et al. [23]). The measure is based on the application of the basic social justice conception that equally needy members of a group should share equally the resources ("the pie") available to the group (see Avi-Itzhak et al. [5]). Accordingly, all customers present in the system at epoch t deserve equal service rate at that epoch, and deviations from that principle result in discrimination (positive or negative). A more extensive review of RAQFM and its adaptation to our models is given in Section 2.2.

Two common mechanisms used in queueing systems to grant preferences to different classes are: a) *Prioritization*, in which the classes are ordered and priority (either preemptive or non-preemptive) is given to higher priority classes over low priority classes, and b) *Resource dedication*, in which each class has a server (or a set of servers) and a queue dedicated to it. Our focus will be on studying these two mechanisms.

In analyzing the fairness aspects of these systems we will first of all be interested in the overall fairness of the system and how the queueing mechanisms affect it. This will be evaluated using the RAQFM measure developed in [23]. Nonetheless, dealing with classes we realize that within this context an additional important quantity is the relative treatment given to the different classes. To this end we introduce a new metrics, called *class discrimination*, which accounts for the expected discrimination experienced by the customers of a certain class. This new metrics is based on the RAQFM measure, and its values are either

positive, negative or zero (reflecting positive discrimination, negative discrimination or no discrimination, respectively). Analysis of this new metrics, its properties and its relations to the RAQFM unfairness measure is given in Section 2.3 and Section 2.4. In particular we show that the (weighted) discrimination of any class is bounded by the square root of the system unfairness.

Next (Section 3), we study class prioritization. Our focus there is to address the issue of how fair it is to grant "full" priority to short jobs, that is, to prioritize all short jobs residing in the system over *all* long jobs residing in the system. There is a large body of literature on such priority schemes (e.g. Avi-Itzhak and Naor [6], Avi-Itzhak [2, 3], Jaiswal [12], Kleinrock [15], Takagi [27]) where the focus in evaluating system performance is on the system expected waiting time, or in a more general framework, the mean waiting cost, under linear cost parameters varying across the classes. Optimization of the system with non-preemptive priorities, based on this performance objective, shows (e.g. Cox and Smith [8, pp. 84-85]) that the optimal scheduling policy is to provide a higher priority to jobs with smaller mean service times (or when costs are involved, apply the μC rule). Such priority may, however, result with long jobs waiting for the completion of many short jobs who arrive behind them, and thus, possibly, to unfair treatment by the system. Thus, system operation that accounts both for efficiency and fairness, might have to resort to a different scheduling.

We start (Section 3.1) with providing a "justification" for short job prioritization. We show that for any service policy that selects customers for service *independently* of their required *service times* (that is, does not "discriminate" based on service time), the discrimination experienced by a customer is monotone non-decreasing in its service time. This means that in such systems an implicit discrimination is applied in favor of the long jobs and against the short jobs. This general result suggests that, from *fairness perspective*, providing preferential service to shorter jobs may be justified in many cases. We then (Section 3.2) study the effect that class prioritization can have on class discrimination. We show that under *general arrival and service conditions*, the class discrimination of the *highest* priority class is always positive, while the class discrimination of the lowest priority class is always negative. Nonetheless, we show that in a multi-class system the class discrimination of a higher priority class is not necessarily higher than that of a lower priority class. We then (Section 3.3) move to provide an analysis of unfairness for an $M/M/1$ type system with two customer classes and class prioritization. The results show that in many cases prioritization of the short jobs over the long jobs leads to higher fairness (than that of FCFS); nonetheless, in some cases FCFS is more fair. We also discuss how the analysis can be extended to the non-exponential service-time case and to a larger number of classes.

In Section 4, we turn to deal with the dedication of resources to classes, where the common strategy is to construct a multi-server system in which a server (or a set of servers) is dedicated to each class. These are very common in human-service facilities, including airport passport control systems (divided to alien and non-alien classes) and public restrooms. Since in some of these systems customers are classified based on personal properties (e.g. gender or nationality) fairness aspects of these systems are highly important. An operational question of interest is whether to allocate equal amount of resources to the different classes or to grant

more resources to the class with the larger service time. The answer to this question is not immediate since one of the basic principles of the RAQFM fairness measure is that short jobs should get preference over long jobs. We first analyze a system consisting of 2 classes, each allocated a single server having the same rate. For a system consisting of two $GI/M/1$ queues we prove that if either i) The inter-arrival time of class 1 is stochastically larger (in dominance sense) than that of class 2, or ii) the mean service time of class 1 is smaller than that of class 2, then class 1 experiences positive class discrimination while class 2 experiences negative class discrimination. While this might sound intuitive it does not hold under all conditions: For systems consisting of arbitrary $G/G/1$ queues and where only the load of class 1 is larger than that of class 2 show that this is not necessarily the case. Second, we deal with how to compute class discrimination in a system with several classes (either several $M/M/1$'s or several $M/GI/1$'s). We propose an algorithm that utilizes the structure of this system and which computes its class discrimination in polynomial time, despite the fact that the state-space complexity of this problem is exponential. We conclude with analyzing the "restroom queueing problem" via an example with two classes of equal arrival pattern and different service times; the results, which might sound somewhat surprising, are that neither allocating the resources equally, nor allocating them *proportionally* to the service times lead to minimization of class discrimination absolute values. Concluding remarks are given in Section 5.

The results of this work can be used in two ways. First, the basic (and general) properties derived, which may sound "intuitive" to many, can be used to build confidence and trust in the fairness measure used. Second, based on this confidence, the more advanced results as well as the fairness evaluation techniques can be used to evaluate the fairness of actual systems and practices.

1.2 Alternative Fairness Measures and Related Work

Several alternative measures for evaluating queue fairness have been proposed in the literature (Sherry-Gordon [25], Avi-Itzhak and Levy [4], Wierman and Harchol-Balter [28], Sandmann [24]). We find those measure less suitable than RAQFM for the analysis carried out in this work for the following reasons: 1) The work in [25] is only an approach for fairness treatment without devising a specific measure; furthermore, it deals only with relative seniority of jobs and not with their sizes. 2) The measure provided in [4] deals mainly with the relative seniority of jobs and is less appropriate to account for different job sizes, 3) The slow-down approach proposed in [28] provides only a fairness *criterion* and not a measure, and 4) The approach proposed in [24] does deal both with seniority and size aspects; however, this work has been only a limited study and it is not clear yet what are the properties of that measure. For a comprehensive overview and comparison of recent proposed approaches and measures of fairness see [5].

2 System Model, the Measures Used and Their Basic Propertoes

2.1 System Model

Consider a queueing system with M servers. Customers are indexed $C_1, C_2, \ldots$, and arrive according to this order. Let a_l and d_l denote the arrival and departure epochs of C_l respectively and let et s_l denote the service requirement (measured in time units) of C_l. Each customer belongs to one of U classes, indexed $1, 2, \ldots, U$. The arrival rate of class u customers is denoted λ_u where $\sum_{u=1}^{U} \lambda_u = \lambda$. An order of priorities is assigned to the classes, where lower class index means higher priority.

We define the *Preemptive Priority* class of scheduling policies. In this class of scheduling policies the priority of a customer is defined to be the priority of the class to which it belongs and the server always serves the highest priority customer present in the system. If a higher priority customer arrives, and finds a lower priority customer in service, the served customer is displaced by the arriving customer. The order of service within each class of customers is usually FCFS. In the *Preemptive Resume* variant, a specific policy analyzed in Section 3.3.2 the preempted customer returns to the head of the queue of its class, and resumes its service from the point it was interrupted, upon reentering service. For discussion of this, and other variants, see Takagi [27, sec. 3.4].

2.2 System Unfairness Measure: RAQFM, And Its Basic Properties

RAQFM was proposed in [23] and defined there for a single server with fixed rate. Below we generalize it to multiple servers with time varying rate. RAQFM evaluates the unfairness in the system as follows: The basic fundamental assumption is that at each epoch, all customers present in the system, deserve an *equal share* of the *total service granted* by the system at that epoch. If we let $0 \leq \omega(t) \leq M$ denote the total service rate granted at epoch t (which often is an integer equaling the number of working servers at that epoch), then the fair share, called the momentary *warranted service* rate, is $\omega(t)/N(t)$, where $N(t)$ is the number of customers in the system at epoch t.

Let $\sigma_l(t)$ be the momentary rate at which service is given to C_l at epoch t. This is called the momentary *granted service* rate of C_l.

The momentary discrimination rate of C_l at epoch t, denoted $c_l(t)$, equals, when C_l is in the system, the difference between its granted service and warranted service,

$$c_l(t) = \sigma_l(t) - \frac{\omega(t)}{N(t)}, \tag{1}$$

and $c_l(t) \stackrel{def}{=} 0$ if C_l is not in the system at epoch t.

The total discrimination of C_l, denoted D_l, is

$$D_l = \int_{a_l}^{d_l} c_l(t)dt. \tag{2}$$

Remark 2.1 (An alternative possible definition of the momentary warranted service and discrimination). The definition of the momentary warranted service (and thus discrimination) given above is based on the concept that a customer deserves an equal share of the *resources granted* by the system at that epoch, $(\omega(t))$. If some of the resources are not granted at epoch t, e.g., due to system idling, or due to the use of only part of the servers, it may be considered as being *inefficient* but not as a discrimination and unfairness.

One could consider an alternative concept by which at epoch t a customer deserves an equal share of *all the available system resources*. Under the notation given above this means that the warranted service will be defined as $m(t)/N(t)$ (instead of $\omega(t)/N(t)$), where $m(t)$ is the total service rate *available* at epoch t (e.g. in a $G/G/M$ type system with M identical servers, $m(t) = M, t \geq 0$). The momentary discrimination given in (1) will be replaced by $c_l(t) = \sigma(t) - m(t)/N(t)$.

The difference between the two alternatives is conceptual and relates to situations where the system does not grant all of its resources. This issue and the differences between the alternatives are more pronounced in multi-server multi-queue systems, and thus is discussed in depth in a study that focuses on these systems (Raz et al. [22]). In this work we choose to focus on the concept of fair division of the *granted* resources (Equation (1)).

For work conserving systems (defined as systems in which the total service given to a customer over time equals its service requirement, i.e. $\int_0^\infty \sigma_l(t) = s_l$), we have from (1) and (2)

$$D_l = s_l - \int_{a_l}^{d_l} \frac{\omega(t)}{N(t)}dt. \tag{3}$$

An important property of RAQFM (shown for a single server, work conserving, and non idling system in [23], and extended here to the measure formulation presented in this work) is the following:

Theorem 2.1 (Zero Expected Discrimination)**.** *In a stationary system, the expected value of discrimination always obeys* $\mathbb{E}\{D\} = 0$.

Proof. Follows immediately from the definition of the momentary discrimination rate that sums to zero at all time epochs. □

As the expected value of discrimination always obeys $\mathbb{E}\{D\} = 0$, according to RAQFM, the unfairness of the system is defined as the variance of the discrimination, which is equal to the second moment, namely $\mathbb{E}\{D^2\}$.

2.3 Class Discrimination and Its Basic Properties

For systems with customer classification it is important to evaluate the treatment given to each class. Such evaluation is mostly important when the classification is based on external customer parameters (e.g. gender, nationality, type of computer application, etc.). Nonetheless, class-based evaluation can be important in many other cases, e.g. when the system does not classify jobs but the operator wants to know whether short jobs are badly discriminated compared to long jobs. Thus, our model and analysis will be general to account for all such cases.

To deal with the relative treatment received by a certain class we introduce the notion of *class discrimination* which relates to the discrimination experienced by a certain class of the population. For class u the discrimination experienced by an arbitrary customer when the system is in steady state is a random variable denoted $D_{(u)} = D|C \in u$. Our interest will be in the expected discrimination experienced by u's customers, namely $\mathbb{E}\{D_{(u)}\}$, termed *Class Discrimination*.

A direct derivation of this value might be difficult. However, in some cases we find the following observation useful. The *momentary discrimination rate* of class u at time t is the sum of discriminations over all u's customers present in the system at time t. Let $\widetilde{D_{(u)}}(t) = \sum_{l \in u} c_l(t)$ denote this variable. Let $\widetilde{D_{(u)}} = \lim_{t\to\infty} \widetilde{D_{(u)}}(t)$ be a random variable denoting the momentary discrimination rate of class u when the system is in steady state. We observe that the relationship between the variables $D_{(u)}$ and $\widetilde{D_{(u)}}$ is analogous to the equilibrium relationship between the variables of *customer delay* (delay experienced by an arbitrary customer) and *number of customers in the system* (number of customers present at an arbitrary moment) in an arbitrary queueing system. We recall from the literature, that often the expected value of the former (customer delay) is of interest but the expected value of the latter is easier to derive. It is in these cases, that Little's Law ([16]), which relates these expected values by the relationship $N = \lambda T$, is used (where T, λ and N denote expected customer delay, arrival rate and expected number in the system, respectively).

Using an analogy to Little's Law (and the same machinery used in its proof) one can now derive a "discrimination version" of Little's Law, namely:

$$\mathbb{E}\{\widetilde{D_{(u)}}\} = \lambda_u \mathbb{E}\{D_{(u)}\}. \tag{4}$$

Remark 2.2. Obviously, this rule can be applied to the whole population (rather than to a certain class), resulting in $\mathbb{E}\{\widetilde{D}\} = 0$.

2.4 The Relation between System Unfairness and Class Discrimination

The following analysis relates the system unfairness (expressed by $\mathbb{E}\{D^2\}$) to the expected class discrimination. We first show that if the overall unfairness is small then so is the absolute value of class discrimination of every class.

Theorem 2.2. *The class discrimination of class u is bounded from above by the overall system unfairness as follows:*

$$\frac{\lambda_u}{\lambda}|\mathbb{E}\{D_{(u)}\}| \leq \sqrt{\mathbb{E}\{D^2\}}.$$

Proof. Since $\mathbb{E}\{{D_{(u)}}^2\} - (\mathbb{E}\{D_{(u)}\})^2 \geq 0$ we have $\frac{\lambda_u}{\lambda}|\mathbb{E}\{D_{(u)}\}| \leq \frac{\lambda_u}{\lambda}\sqrt{\mathbb{E}\{{D_{(u)}}^2\}}$. But

$$\frac{\lambda_u}{\lambda}\sqrt{\mathbb{E}\{{D_{(u)}}^2\}} \leq \sqrt{\frac{\lambda_u}{\lambda}\mathbb{E}\{{D_{(u)}}^2\}} \leq \sqrt{\sum_{i=1}^{U}\frac{\lambda_i}{\lambda}\mathbb{E}\{{D_{(i)}}^2\}} = \sqrt{\mathbb{E}\{D^2\}}.$$

□

Corollary 2.1. *Consider an arbitrary system with U customer classes. If the system unfairness obeys $\mathbb{E}\{D^2\} = 0$ then for every class $1 \leq u \leq U$ the class discrimination obeys $\mathbb{E}\{D_{(u)}\} = 0$.*

The proof is immediate from Theorem 2.2.

Theorem 2.3. *Consider a system with U classes. Assume that the class discrimination of each class u obeys $\mathbb{E}\{D_{(u)}\} = 0$. Then the system unfairness, $\mathbb{E}\{D^2\}$ can still be positive.*

Proof. (by example). Consider a system with two classes, A and B. Assume that the service requirement is one unit for all customers and the arrival process is in pairs, one customer of each type. Assume that the inter-arrival time, between two consecutive arrival epochs, is given by $x > 2$ and that the server serves half of the pairs in the order A first B last, and half of them in reverse order. One can easily observe that half of the customers experience positive discrimination of 0.5 and half experience negative discrimination of -0.5. Thus $\mathbb{E}\{D^2\} = 0.25$. Nonetheless the expected class discrimination is zero for both classes. □

The implications of these results are as follows: 1) If one maintains very low system unfairness it guarantees that the class discrimination of large population classes (classes with relatively high arrival rates) will be very small; the discrimination of a lightly populated class can still be very high. 2) Maintaining low class discrimination to all classes does not guarantee a fair system, since there could be unfairness in treatment of customers within a class.

3 Class Prioritization

In this section we study the effect class prioritization has on the system unfairness. We focus on studying the common practice of prioritizing short jobs. We first show that generally speaking, prioritizing short jobs is justified, since otherwise these jobs are negatively discriminated. We then show the effectiveness of class prioritization. We show that while prioritization can guarantee positive discrimination to the class with highest priority and negative discrimination to the class with lowest priority, it cannot guarantee monotonicity in discrimination.

3.1 Prioritizing Short Jobs is Jusitified

As mentioned above, we start by showing one justification why prioritizing short jobs is justified, from the discrimination point of view.

Definition 3.1 (Stochastic Dominance Between Random Variables)**.** Consider non negative random variables X_1, X_2 whose distributions are $F_{X_1}(t) = \mathbb{P}\{X_1 \leq t\}, F_{X_2}(t) = \mathbb{P}\{X_2 \leq t\}$. We say that X_1 *stochastically dominates* X_2, denoted $X_1 \succ X_2$, if $F_{X_1}(t) \leq F_{X_2}(t) \quad \forall t \geq 0$.

Theorem 3.1. *Let C_l be a customer with service requirement s_l. Consider a $G/G/M$ system under non-preemptive service policy, where the service decision is independent of the service times. Let $D_l^{(s_l)}$ be a random variable denoting the discrimination of C_l, when it arrives at the system in steady state. Then $D_l^{(s_l)}$ is monotone non-decreasing in s_l, namely if $s_l' > s_l$ then $D_l^{(s_l')} \succ D_l^{(s_l)}$.*

Proof. Consider service times $s_l, s_l', \quad s_l' > s_l$. Observe a customer C_l. Under any non-preemptive service policy, C_l waits until epoch q_l, when it enters service, and stays in service until its departure. (2) can thus be written as

$$D_l = \int_{a_l}^{q_l} c_l(t)dt + \int_{q_l}^{d_l} c_l(t)dt. \tag{5}$$

The first term in this sum is independent of the service requirement. In the second term $d_l - q_l = s_l$.

To prove the monotonicity we consider a specific sample path π and compare the values of $D_l^{(s_l)}$ and $D_l^{(s_l')}$ for this path, denoted by $D_{l,\pi}^{(s_l)}$ and $D_{l,\pi}^{(s_l')}$. From (5) we have

$$D_{l,\pi}^{(s_l')} - D_{l,\pi}^{(s_l)} = \int_{q_l+s}^{q_l+s'} c_l(t)dt \geq 0, \tag{6}$$

where the inequality is due to $c_l(t) \geq 0$, which follows from (1). Since (6) holds for every sample path π, the proof follows. □

Theorem 3.2. *Let S_l and S_l' be random variables representing two alternate service times of C_l and let $F_{S_l}(t)$ and $F_{S_l'}(t)$ be their distribution functions. Consider a $G/G/M$ system under non-preemptive service policy, where the service decision is independent of the service times. If $S_l' \succ S_l$ then $D^{(S_l)} \succ D^{(S_l')}$.*

Proof. The proof follows by applying Theorem 3.1 for the whole range of service times, and using the fact that $S_l' \succ S_l$. □

Corollary 3.1. *Consider a $G/G/M$ system under non-preemptive service policy, where the service decision is independent of the service times. Let $D^{(S_l)}$ be a random variable denoting the discrimination of C_l, when it arrives at the system in steady state, given its service time S_l (random variable). Then $\mathbb{E}\{D^{(S_l)}\}$ is monotone non-decreasing in S_l. That is, if $S_l' \succ S_l$ then $\mathbb{E}\{D^{(S_l')}\} \geq \mathbb{E}\{D^{(S_l)}\}$.*

Similarly, using class notation, let S_u be the service requirement distribution of class u customers. Then $S_u \succ S_{u'} \Rightarrow \mathbb{E}\{D_{(u)}\} \geq \mathbb{E}\{D_{(u')}\}$.

Remark 3.1. Using the same arguments it can be shown that Theorem 3.1 also holds in the case of a preemptive system, providing that the preemption of a customer with service $s' > s$, during the period at which it receives the first s units of service, is unchanged, i.e. preemptions are not determined by the length of the service required by the customer. In a similar manner Theorem 3.2 will hold as well.

In conclusion, we have shown that service policies that do not give preferred service to shorter jobs, actually discriminate against those jobs. This provides one more justification for prioritizing shorter jobs.

3.2 The Effect of Class Prioritization

We now move on to study how class prioritization affects the class discrimination (recall that $\mathbb{E}\{D_{(u)}\}$ denotes the class discrimination for class u, as defined in Section 2.3).

Theorem 3.3. . *In a $G/G/M$ system, with U classes, if the scheduling policy belongs to the class of preemptive priority scheduling policies, then $\mathbb{E}\{D_{(1)}\} \geq 0$ and $\mathbb{E}\{D_{(U)}\} \leq 0$.*

Proof. Let $N_u(t)$ be the number of class u customers in the system at epoch t. As the scheduling policy belongs to the class of preemptive priority scheduling policies, if $N_1(t) \leq M$, then all $N_1(t)$ customers are served at epoch t. Otherwise, M out of them are served. Thus

$$\widetilde{D_{(1)}}(t) = \begin{cases} N_1(t) - \frac{\omega(t)N_1(t)}{N(t)} & N_1(t) \leq M \\ M - \frac{MN_1(t)}{N(t)} & N_1(t) > M \end{cases} = \begin{cases} N_1(t)\left(1 - \frac{\omega(t)}{N(t)}\right) & N_1(t) \leq M \\ M\left(1 - \frac{N_1(t)}{N(t)}\right) & N_1(t) > M \end{cases}, \tag{7}$$

which is greater or equal to zero since $\omega(t) \leq N(t)$ and $N_1(t) \leq N(t)$.

Thus, $\widetilde{D_{(1)}}(t) \geq 0 \Rightarrow \widetilde{D_{(1)}} \geq 0 \Rightarrow \mathbb{E}\{\widetilde{D_{(1)}}\} \geq 0$, and from (4), $\mathbb{E}\{D_{(1)}\} \geq 0$.

Note that (7) also provides the only epochs in which $\widetilde{D_{(1)}}(t) = 0$, namely when either $N_1(t) = N(t)$ (all the customers in the system are of class 1), or $N(t) < M$ (there are less than M customers in the system), or $N_1(t) = 0$. In fact, for every class u, $\widetilde{D_{(u)}}(t) = 0$ when either $N_u(t) = N(t)$, or $N(t) < M$, or $N_u(t) = 0$.

As for $\widetilde{D_{(U)}}(t)$, it equals zero when either $N_U(t) = N(t)$, or $N(t) < M$, or $N_U(t) = 0$. Otherwise there are two cases, either $N(t) - N_U(t) \geq M$ or $N(t) - N_U(t) < M$. In the first case there are more than M customers of higher priority in the system, and thus no class U customers are being served. Therefore, $\widetilde{D_{(U)}}(t) = -N_U(t)M/N(t)$ which is negative. In the second case there are some class U customers being served. In this case let $\omega_U(t)$ be the number of class U customers served at epoch t. Using this notation

$$\widetilde{D_{(U)}}(t) = \omega_U(t) - \frac{N_U(t)M}{N(t)} = \frac{\omega_U(t)N(t) - N_U(t)M}{N(t)}. \tag{8}$$

To prove that this value is negative, let $N'(t) = N(t) - M$ denote the number of customers waiting at epoch t, all of whom must be of class U. We can write $N(t) = M + N'(t)$,

$N_U(t) = \omega_U(t) + N'(t)$. Substituting into (8) yields

$$\widetilde{D_{(U)}}(t) = \frac{\omega_U(t)(M + N'(t)) - (\omega_U(t) + N'(t))M}{N(t)} = \frac{(\omega_U(t) - M)N'(t)}{N(t)} < 0,$$

since $\omega_U(t) < M$. Thus, $\widetilde{D_{(U)}}(t) \leq 0 \Rightarrow \widetilde{D_{(U)}} \leq 0 \Rightarrow \mathbb{E}\{\widetilde{D_{(U)}}\}$, and from (4), $\mathbb{E}\{D_{(U)}\} \leq 0$. □

The striking thing about Theorem 3.3 is that in the context of a 2 class system one may conclude from Corollary 3.1 that to prevent negative discrimination of the short job class one should prioritize it over the long job class. However, as Theorem 3.3 shows, this will not only reduce the class discrimination, but also reverse it, leading to positive discrimination of the short job class, regardless of how small they are.

One interesting case is the case where there are two classes, one of which has infinitesimal service requirements. The following theorem shows that in that case the class discrimination of both class tends to zero.

Theorem 3.4. *In a $G/G/M$ system, with 2 classes, where the scheduling policy belongs to the class of preemptive priority scheduling policies, if the mean service time $1/\mu_1 \to 0$, and $\lambda_1/\lambda_2 \nrightarrow 0$ then $\mathbb{E}\{D_{(1)}\} \to 0, \mathbb{E}\{D_{(2)}\} \to 0$.*

Proof. First note that in any general system the following simple conservation law holds:

$$\frac{\sum_{u=1}^{U} \lambda_i \mathbb{E}\{D_{(i)}\}}{\sum_{u=1}^{U} \lambda_i} = 0. \tag{9}$$

This law is simple to derive and results from the fact that $\mathbb{E}\{D\} = 0$. It is interesting to draw the similarity of this law to the pseudo-conservation law regarding the waiting time of customer classes in a single server systems (see Kleinrock [15, Chap. 3.4]).

According to Theorem 3.3, $\mathbb{E}\{D_{(1)}\} \geq 0$. Observe that in (3) the only positive part is s_l and thus $D_l < s_l$. Taking expectations on both sides, and conditioning on class 1 customers we get $0 \leq \mathbb{E}\{D_{(1)}\} \leq 1/\mu_1$, and since both the upper and the lower limits tend to zero $\mathbb{E}\{D_{(1)}\} \to 0$. Using (9) we get $\mathbb{E}\{D_{(2)}\} = \lambda_1/\lambda_2 \mathbb{E}\{D_{(1)}\}$ and since $\lambda_1/\lambda_2 \nrightarrow 0$ this means $\mathbb{E}\{D_{(2)}\} \to 0$. □

Remark 3.2. If $\lambda_1 \gg \lambda_2$ then class 2 might have negative discrimination. This corresponds to a case where there are very few large customers, swamped by a multitude of smaller, prioritized, customers.

Having shown that discrimination of the most prioritized class is always positive, and that of the least prioritized class is always negative, one might expoect that the discrimination is monotonic with the class priority. However, as the following example shows, this is not the case. Consider a 4-class $M/M/1$ type system with preemptive resume priority. All four classes have an arrival rate of 0.01, and all but class 2 have a mean service requirement of 10 ($\mu = 0.1$). For class 2 we will consider $\mu = 0.1, 0.2, 0.3, 0.4, 0.5$. Figure 1 depicts the class

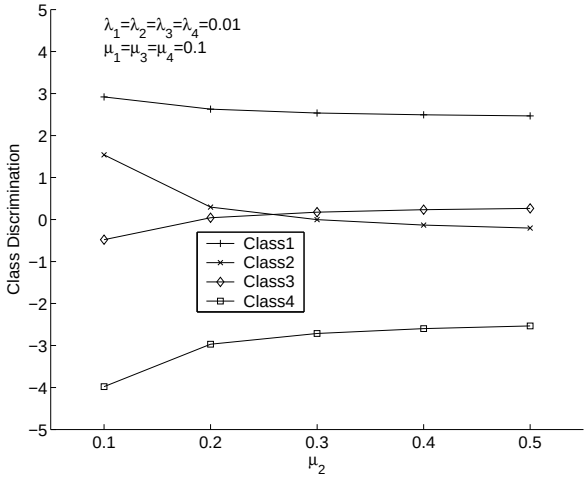

Figure 1: The Effect of Preemptive Priority on Four Classes of Customers

discrimination for the four classes. The results were achieved through simulation, although, as shown in the next section, similar results can be achieved through numerical analysis.

Observe that when the service requirement of class 2 is equal to that of the other classes, the class discrimination is monotonic with the class priority. However, when class 2 has smaller service requirement, this is not the case. This means that class prioritization is limited in its effect in multiple class systems.

3.3 Fairness Analysis of Preemptive Resume Priority Scheduling in Single Server Systems

In this section we provide the analysis of class prioritization in systems with a single server and Poisson arrivals. The analysis is given in detail for exponential service times and two customer classes (Sections 3.3.1, 3.3.2 and 3.3.3) and an explanation is given how it can be generalized to arbitrary phase-type service times (Section 3.3.4), which can approximate general service times ($M/GI/1$ type system), or to larger number of classes. The analysis is based on the methodology developed in [23] which is extended here to deal with classes and prioritization. Our analysis starts (Section 3.3.1) with some definitions and simple observations which are useful in general for the analysis of discrimination and fairness in $M/M/1$ type system, and then moves to the specific systems we would like to analyze.

3.3.1 Analysis of Fairness and Discrimination in $M/M/1$ type systems

Consider a $M/M/1$ type system, with U classes of customers, where class u arrivals follow a Poisson process with rate λ_u, and their required service times are i.i.d. exponentially with mean $1/\mu_u$, $u = 1, 2, \ldots, U$. The total arrival rate is denoted by $\lambda \stackrel{def}{=} \sum_{u=1}^{U} \lambda_u$ and, for stability, it is assumed that $\rho \stackrel{def}{=} \sum_{u=1}^{U} \lambda_u/\mu_u < 1$.

The analysis is based on the observation that in $M/M/\cdot$ systems, time can be viewed as slotted by arrival and departure epochs. We will limit our analysis to systems where service decisions are made only on arrival and departure epochs. Thus, the number of customers and the rate of service given to each customer are constant during each slot, and thus so is the momentary discrimination rate.

In more formal terms, arrival and departure epochs are labeled event epochs, and time is viewed as being slotted by these event epochs. The i-th time slot, of duration $T_i, i = 1, 2, \ldots,$ is bounded by the $(i-1)$-th and the i-th event epochs. We define $0 \leq \omega_i \leq u$ as the total service rate given in the i-th slot, $\sigma_{i,l}$ as the rate at which service is given to C_l at the i-th slot, and N_i as the number of customers in the system during the i-th slot. Using these, $c_{i,l}$, the momentary discrimination of C_l at the i-th slot is

$$c_{i,l} = \sigma_{i,l} - \frac{\omega_i}{N_i}. \tag{10}$$

In this formulation we modify a_l, d_l to denote the indexes of the arrival and departure slots of C_l respectively (C_l arrives at the beginning of the a_l-th slot and departs at the end of the d_l-th slot). The total discrimination accumulated for C_l during the i-th slot is $c_{i,l}T_i$. Thus, the slotted version of (2) is

$$D_l = \sum_{i=a_l}^{d_l} c_{i,l} T_i,$$

In the specific service policies we are going to analyze, service is given to one customer at a time. Assume a class u customer, $u = 1, \ldots, U$, is served in a given slot. The first two moments of that slot's duration, $t_u^{(1)}$ and $t_u^{(2)}$, are

$$t_u^{(1)} = \frac{1}{\lambda + \mu_u}, \qquad t_j^{(2)} = \frac{2}{(\lambda + \mu_u)^2} = 2(t_u^{(1)})^2. \tag{11}$$

The probability that a slot in which a class u customer is served, ends with an arrival of a class k customer, is denoted $\tilde{\lambda}_{u,k}$. The probability that a slot in which a class u customer is being served, ends with an arrival of any customer, is denoted $\tilde{\lambda}_u$. The probability that a slot where a class j customer is being served, ends with the departure of the same customer, is denoted $\tilde{\mu}_u$. We have:

$$\tilde{\lambda}_{u,k} = \frac{\lambda_k}{\lambda + \mu_u}, \qquad \tilde{\lambda}_u = \frac{\lambda}{\lambda + \mu_u}, \qquad \tilde{\mu}_u = \frac{\mu_u}{\lambda + \mu_u}. \tag{12}$$

3.3.2 Priority Scheduling for Two Customer Classes

We now move on to analyze the specific system we are interested in, namely a full priority system with two classes. The discipline analyzed here is the preemptive resume one, as defined in Section 1.

The steps we take in the analysis are as following. (i) First we define a state space in which the analysis will be made and express the momentary discrimination as a function of

the state variables. (ii) Then we express the class discrimination and the unfairness as a function of the discrimination of a customer, given that a certain number of customers of any type was seen by that customer on arrival, and connect this with the state space defined in (i). (iii) Lastly we show how to calculate these conditional discriminations. For class 1 customers, we resort to the work in [23]. For class 2 customers we provide sets of linear equations by considering the possible state of the customer at the end of the slot.

(i) Consider an arbitrary tagged customer of class j, denoted C. Let a_i be the number of class i customers ahead of C in the queue. Note that for $j = 1$, $a_2 = 0$, due to the priority class 1 gets over class 2. Let b be the number of customers behind C. Note that for $j = 1$ this includes both class 1 customers behind C in its queue and all class 2 customers in the system, while for $j = 2$ it only includes class 2 customers behind C. Due to the memoryless properties of the system, The state (a_1, a_2, b), denoted $\mathcal{S}_{a_1,a_2,b}$, captures all that is needed to predict the future discrimination of C.

From (10), the momentary discrimination during a slot where C is in state $S_{a_1,a_2,b}$ denoted $c_j(a_1, a_2, b)$, is

$$c_1(a_1,a_2,b) = \begin{cases} -\frac{1}{a_1+b+1} & a_1 > 0 \\ 1 - \frac{1}{b+1} & a_1 = 0 \end{cases}$$
$$c_2(a_1,a_2,b) = \begin{cases} -\frac{1}{a_1+a_2+b+1} & a_1 > 0 \text{ or } a_2 > 0 \\ 1 - \frac{1}{b+1} & a_1, a_2 = 0 \end{cases}.$$

Let $\mathbb{E}\{D_j|k_1, k_2\}$, $j = 1, 2$, denote the expected value of discrimination of a class j customer, given that the customer sees k_1 customers of class 1, and k_2 customers of class 2 on arrival.

(ii) Let P_{k_1,k_2} be the steady state probability that there are k_1 customers of class 1 and k_2 customers of class 2 in the system.

The class discrimination is given as $\mathbb{E}\{D_j\} = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\mathbb{E}\{D_j|k_1,k_2\}P_{k_1,k_2}$. The second moment of class discrimination is $\mathbb{E}\{D_j^2\} = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\mathbb{E}\{D_j^2|k_1,k_2\}P_{k_1,k_2}$ and the unfairness can be expressed as $\mathbb{E}\{D^2\} = \frac{1}{\lambda}\left(\lambda_1\mathbb{E}\{D_1^2\} + \lambda_2\mathbb{E}\{D_2^2\}\right)$.

Numerically calculating the steady state probabilities $P_{i,j}$ can be done, for example, from the balance equations

$$\begin{cases} (\lambda_1+\lambda_2+\mu_1)P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + \mu_1 P_{i+1,j} & i,j > 0 \\ (\lambda_1+\lambda_2+\mu_1)P_{i,j} = \lambda_1 P_{i-1,j} + \mu_1 P_{i+1,j} & i > 0, j = 0 \\ (\lambda_1+\lambda_2+\mu_2)P_{i,j} = \lambda_2 P_{i,j-1} + \mu_1 P_{i+1,j} + \mu_2 P_{i,j+1} & i = 0, j > 0 \\ (\lambda_1+\lambda_2)P_{i,j} = \mu_1 P_{i+1,j} + \mu_2 P_{i,j+1} & i,j = 0, \end{cases} \tag{13}$$
$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P_{i,j} = 1,$$

to any required accuracy.

Remark 3.3. Steady state probabilities of multiple class systems were recently studied in Sleptchenko [26], Harten and Sleptchenko [10], Harten et al. [11] and more. It is not the purpose of this paper to introduce additional results, nor to demonstrate alternative methods to achieve the same results. The reader may use the results in Sleptchenko [26], Harten and Sleptchenko [10], Harten et al. [11] as appropriate. Alternately, one may possibly want to use the Laplace-Stieltjes transforms derived for $M/M/c$ and $M/G/1$ Vacations, with priorities, see Kella and Yechiali [13, 14].

Let $D_j(a_1, a_2, b)$ be a random variable, denoting the discrimination experienced by a class j customer, through a walk starting at $\mathcal{S}_{a_1,a_2,b}$, and ending at its departure. Then

$$\mathbb{E}\{D_1|k_1, k_2\} = \mathbb{E}\{D_1(k_1, 0, k_2)\}$$
$$\mathbb{E}\{D_2|k_1, k_2\} = \mathbb{E}\{D_2(k_1, k_2, 0)\}.$$

(iii) Let $d_j(a_1, a_2, b)$ and $d_j^{(2)}(a_1, a_2, b)$ be the first two moments of $D_j(a_1, a_2, b)$.

We first analyze $d_1(a_1, a_2, b)$. Note that a class 1 customer sees the system as a single class FCFS queue. As the customer arrives, all class 1 customers in the system are in front of him in the queue, and class 2 customers are behind him (as opposed to a single class system where all customers are in front of an arriving customer). However, from that moment on, the types of customers behind the customer do not matter to the calculation, and thus the system behaves like a queue with an arrival rate of λ. All departures are of class 1 customers, and therefore the departure rate is μ_1.

The single class FCFS system was analyzed in [23]. Using the notation in that paper, $\mathcal{S}_{a,b}$ represents the state where there are a customers ahead of C and b customers behind him. Thus, as analyzed there, a customer arriving into a single class FCFS system, seeing k customers in the system on arrival, starts his walk at $\mathcal{S}_{k,0}$. Similarity, a class 1 customer in a system with two classes, seeing k_1 customers of class 1, and k_2 customers of class 2 on arrival, starts his walk at $\mathcal{S}_{k_1,k_2}$. Thus, we resort to the analysis of the single class FCFS system.

Let $D(a, b)$ be a random variable, denoting the discrimination experienced by a customer, through a walk starting at $\mathcal{S}_{a,b}$, and ending at its departure, with first two moments $d(a, b)$ and $d^{(2)}(a, b)$. This leads to

$$\mathbb{E}\{D_1|k_1, k_2\} = \mathbb{E}\{D(k_1, k_2)\} = d(k_1, k_2)$$
$$\mathbb{E}\{D_1^2|k_1, k_2\} = \mathbb{E}\{D(k_1, k_2)^2\} = d^{(2)}(k_1, k_2).$$

Expressions for deriving $d(a, b)$ and $d^{(2)}(a, b)$ were given in [23].

For a customer of class 2, assume C is in state $\mathcal{S}_{a_1,a_2,b}$ at slot i. Let u be the type of customer being served in this slot. Then u equals 1 when $a_1 > 0$ and 2 when $a_1 = 0$, i.e. u is implied directly from a_1 and thus does not have to be accounted for in the state.

The slot length is exponentially distributed with first two moments $t_u^{(1)}$ and $t_u^{(2)}$ (from (11)). At the slot end, the system will encounter one of the following events and C's state will change accordingly:

1. A class 1 customer arrives at the system. The probability of this event is $\tilde{\lambda}_{u,1}$. C's state changes to $\mathcal{S}_{a_1+1,a_2,b}$.

2. A class 2 customer arrives at the system. The probability of this event is $\tilde{\lambda}_{u,2}$. C's state changes to $\mathcal{S}_{a_1,a_2,b+1}$.

3. A customer leaves the system. The probability of this event is $\tilde{\mu}_u$. If C is being served ($a_1 = 0, a_2 = 0$), C leaves the system. Otherwise, if $a_1 > 0$, C's state changes to $\mathcal{S}_{a_1-1,a_2,b}$, and if $a_1 = 0$, C's state changes to $\mathcal{S}_{0,a_2-1,b}$.

All probabilities are from (12). This leads to the following recursive expressions

$$d_2(a_1,a_2,b) = \begin{cases} t_u^{(1)} c_2(a_1,a_2,b) + \\ \quad \tilde{\lambda}_{u,1} d_2(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2(a_1,a_2,b+1) + \tilde{\mu}_u d_2(a_1-1,a_2,b) & a_1 > 0 \\ t_u^{(1)} c_2(a_1,a_2,b) + \\ \quad \tilde{\lambda}_{u,1} d_2(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2(a_1,a_2,b+1) + \tilde{\mu}_u d_2(a_1,a_2-1,b) & a_1 = 0, a_2 > 0 \\ t_u^{(1)} c_2(a_1,a_2,b) + \tilde{\lambda}_{u,1} d_2(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2(a_1,a_2,b+1) & a_1 = 0, a_2 = 0 \end{cases} . \tag{14}$$

$$d_2^{(2)}(a_1,a_2,b) = \begin{cases} t_u^{(2)} (c_2(a_1,a_2,b))^2 + \tilde{\lambda}_{u,1} d_2^{(2)}(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2^{(2)}(a_1,a_2,b+1) + \\ \quad \tilde{\mu}_u d_2^{(2)}(a_1-1,a_2,b) + 2t_u^{(1)} c_2(a_1,a_2,b) & a_1 > 0 \\ \quad \left(\tilde{\lambda}_{u,1} d_2(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2(a_1,a_2,b+1) + \tilde{\mu}_u d_2(a_1-1,a_2,b) \right) \\ t_u^{(2)} (c_2(a_1,a_2,b))^2 + \tilde{\lambda}_{u,1} d_2^{(2)}(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2^{(2)}(a_1,a_2,b+1) + \\ \quad \tilde{\mu}_u d_2^{(2)}(a_1,a_2-1,b) + 2t_u^{(1)} c_2(a_1,a_2,b) & a_1 = 0, a_2 > 0 \\ \quad \left(\tilde{\lambda}_{u,1} d_2(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2(a_1,a_2,b+1) + \tilde{\mu}_u d_2(a_1,a_2-1,b) \right) \\ t_u^{(2)} (c_2(a_1,a_2,b))^2 + \tilde{\lambda}_{u,1} d_2^{(2)}(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2^{(2)}(a_1,a_2,b+1) & \\ \quad + 2t_u^{(1)} c_2(a_1,a_2,b) \left(\tilde{\lambda}_{u,1} d_2(a_1+1,a_2,b) + \tilde{\lambda}_{u,2} d_2(a_1,a_2,b+1) \right) & a_1 = 0, a_2 = 0 \end{cases} . \tag{15}$$

Remark 3.4 (Extending the Analysis to the Multiple Classes Case). The analysis above can be easily extended to the multiple class ($U > 2$) case, at the expense of higher computational complexity or the use of some approximation. Such an extension is provided in the technical report [21], along with a discussion of the computational repercussions.

3.3.3 Computational Aspects

(13) provides a set of linear equations for calculating $P_{i,j}$. To solve it, one first needs to decide on a maximum relevant number of customers that can be seen on arrival, say K.

This leads to a sparse set of K^2 linear equations. See, for example, Anderson et al. [1], Davis [9], for efficient methods of solving such a set.

Since the evaluation of $\mathbb{E}\{D_j\}$ and $\mathbb{E}\{D_j^2\}$ for $j = 1$ is much simpler than that for $j = 2$ the overall computation is dominated by that of $j = 2$.

(14) and (15) provide a recursive method for calculating $d_2(a_1, a_2, b)$ and $d_2^{(2)}(a_1, a_2, b)$. To solve them one may iterate the equality starting with an initial guess, say zero, until the required relative accuracy is reached. For example, one can keep iterating until the relative change in the sum of absolute values is smaller than some small constant α, say 10^{-10}. If I iterations are needed to reach the required relative accuracy, the time complexity is $O(IK^3)$. As the computation requires keeping a copy of the values of $d_2(a_1, a_2, b)$ for one iteration, the space complexity is $O(K^3)$.

Our experience shows that for $\rho = 0.8$, when choosing $K = 30, \alpha = 10^{-10}$ led to $I = 200$ at the worst case. The results achieved were sufficiently close to those reached by simulation.

Remark 3.5 (The size of K). One way to approximate the size of K is to note that in a single class system, p_n, the steady state probability of having n customers in the system, is $p_n = (1 - \rho)\rho^n$ and the probability of having more than K customers in the system is ρ^{K+1}. The selection of K may be determined by the value one wants to allow for ρ^{K+1}. Suppose $\rho^{K+1} = 10^{-3} \Rightarrow K = -3/\log\rho - 1$, which for $\rho = 0.8$ yields $K = 30$. While this is only an approximation of the system analyzed, practice shows that it is not a bad estimate.

3.3.4 Analyzing Systems with Non-exponential Service Times

The analysis provided above for exponential service times can be extended and applied to systems where the service time is not exponential. In such cases one can model (or approximate) the service time distribution via a Coxian/Erlangian type distribution. In Brosh et al. [7] such an approximation is proposed and studied for single server systems with a single customer class. The technique used in [7] can be readily applied to the single server system with multiple classes of customers.

3.3.5 System Evaluation: Numerical Results

One specific case of interest is the following: suppose that a clerk issues two types of documents, one requiring only his signature, and one requiring his full attention for several minutes. It is common to suggest, due to fairness reasons, that customers with shorter service requirement (those requiring only a signature) should be served ahead of other customers. For simplicity assume that the rates of arrival of both customer classes are equal.

It might be true that this suggestion is indeed fair. This, however, may depend on the parameters, and it is reasonable to predict that the shorter the service times of the priority class are, the greater are the fairness benefits (relative to FCFS). One can therefore predict that there is some minimum ratio of the mean service requirement of the "preferred" class, to that of the rest of the population, below which the priority schedule is more fair than FCFS.

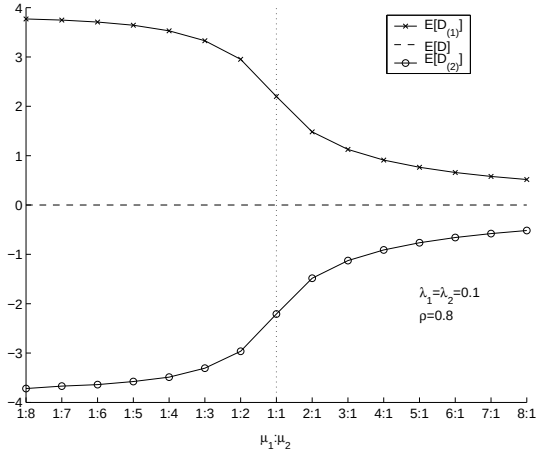

Figure 2: Class Discrimination in a Single Server Priority System with Two Customer Classes

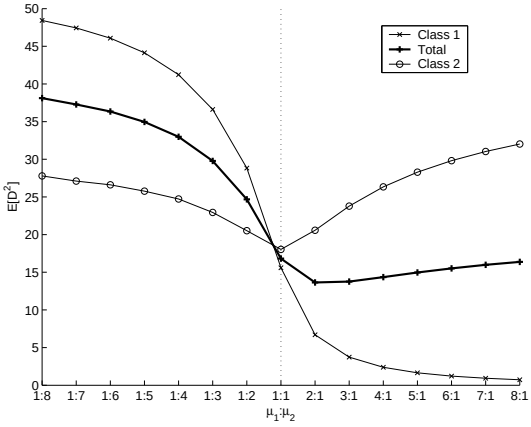

Figure 3: Unfairness in a Single Server Priority system with Two Customer Classes

To this end, we examine the class discrimination and the unfairness as functions of the service difference between the customer classes, expressed by the mean service time ratio μ_1/μ_2. As demonstrated in [23], the unfairness of the system is sensitive to the utilization ρ. Therefore, we maintain constant utilization $\rho = 0.8$, independently of μ_1/μ_2. For simplicity, the evaluation is done for equal arrival rates of $\lambda_1 = \lambda_2 = 0.1$. Figure 2 depicts the class discrimination for the two classes as well as the expected discrimination for the entire system. We observe the following properties:

1. Class 1 is always positively discriminated, and class 2 is always negatively discriminated. Indeed, the discrimination of class 1 customers is at the expense of class 2 customers, since $\mathbb{E}\{D\} = 0$. This result is in fact correct for a much wider set of arrival and service distributions, and for any number of servers, due to Theorem 3.3.

2. The positive (negative) discrimination is monotone-increasing (decreasing) in the expected service requirement of class 1 customers, as expected from Theorem 3.4.

Figure 3 depicts the unfairness, as measured by the second moment of the discrimination, for the two classes, and for the system. We observe the following properties:

1. The highest system unfairness is observed at the left part of the figure. This is the case where very long jobs (class 1) receive priority over the short jobs. This behavior is naturally expected.

2. One may observe that in the right side of the figure, where the shorter jobs receive priority, the system unfairness slightly increases with the service requirement ratio. This might be counter-intuitive at first sight, since priority is given to the short jobs. This increase is however explainable, and results from unfairness between class 2 customers

and themselves, which increases at this region due to the increased variability in service time. Note that the unfairness observed by class 1 customers approaches zero at this range, as expected (see Theorem 3.4, and observe that the same proof applies for the second moment).

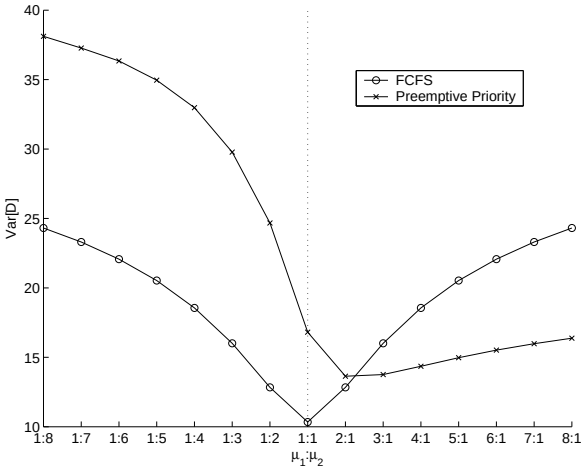

Figure 4: Unfairness in a Single Server Priority System with Two Customer Classes

Finally, Figure 4 compares the unfairness in the FCFS schedule that in the priority one. The analysis of the FCFS schedule was done a similar way. Full details of that analysis can be found in the a technical report [21]. We observe the following properties:

1. When class 1 customers have longer expected service requirement it is less fair, system-wise, to give them priority.

2. When class 1 customers have shorter expected service requirement and the ratio is over $2:1$ it is more fair, system-wise, to use two queues and give priority to the shorter jobs.

To conclude this section, we observe that in the specific case analyzed, there is a threshold value for the ratio of the mean non-priority job size, to the mean prioritized job size. If the ratio is below this threshold, it is more fair to serve the customers in FCFS manner. If the ratio is above this threshold, the priority manner is more fair. We conjecture, and leave it open in the framework of this paper, that this property will also apply to non exponential distributions of service and inter-arrival times.

Recall the clerk case, presented in the beginning of this section. The results seem to agree with common intuition - it is less fair to prioritize a specific class of customers (over another class), unless the service requirement of the prioritized customers, is small enough compared to the others.

4 Resource Dedication to Classes

In this section we deal with the dedication of resources to classes and in evaluating the fairness of this practice. Perhaps the most common approach for dedicating resources to classes is by assigning each class a set of one or more servers and associating each class with a single FIFO queue. A simplified model of these systems is a model where each class receives a dedicated server, whose speed can be used to approximate the number of servers dedicated to the class.

These systems are very common in practice, e.g., customs queues at airports and public restroom queues, and may sometimes cause much frustration to customers. An open question is whether a class with larger service times deserves more resources. This question, in the context of the public restroom system, has been debated publicly and even received attention by some legislative bodies. Nonetheless, this question was not analyzed in a queueing fairness framework. One may offer an immediate "intuitive" positive answer; However, such an answer can be counteracted by a counter "intuitive" argument claiming that shorter jobs should receive preference over longer jobs. The answer to the question is therefore not immediate.

The issue in focus is, therefore, that of class discrimination under resource dedication settings. We will consider a system consisting of U classes, each equipped with its dedicated servers and each served in FIFO order within the class. In Section 4.1 below we will focus on systems with two classes, where each class has a single dedicated server and show that the answer to the question posed above is indeed positive (since under equal resource allocation the more heavily loaded class is subject to negative discrimination) under a wide set of conditions; We further show that this does not always hold and there are cases where the heavily loaded class is positively discriminated. Then in Section 4.2 we consider a system consisting of U classes and provide an algorithmic approach for deriving the class discrimination for it; the algorithm we propose exploits the structure of the problem and yields results in polynomial complexity despite the exponential size of the state space. Lastly, in Section 4.3 we use a numerical example to address the practical question of how to assign servers to classes to reduce class discrimination.

4.1 Dominance Results for 2 Class Systems

We start with showing dominance results for the $GI/M/1$ case.

Theorem 4.1. *Consider a dual class system where each class behaves like a $GI/M/1$ system. Let A_u, S_u, $u = 1, 2$ be random variables denoting the interarrival time and the service requirement, respectively, of class u customers. Then, if either (i) $A_1 \prec A_2$ and $1/\mu_1 \geq 1/\mu_2$, or (ii) $A_1 \preceq A_2$ and $1/\mu_1 > 1/\mu_2$ then $\mathbb{E}\{D_{(1)}\} < \mathbb{E}\{D_{(2)}\}$.*

Proof. Let $\widetilde{D_{(u)}}$ be a random variable denoting the total momentary discrimination rate to class u customers at steady state. $\widetilde{D_{(1)}}$ can be derived by conditioning on the system state and examining several cases: 1) Case 1 - server 1 is idle: In this case no class 1 customers are

present in the system and thus $\widetilde{D_{(1)}} = 0$. 2) Case 2 - server 2 is idle and server 1 is busy: In this case the total warranted service to class 1 customers is 1 and the granted service to the class is also 1. Thus $\widetilde{D_{(1)}} = 0$. 3) Case 3 - server 1 and server 2 are busy: Let $n_i > 0$ be the number of customers present at the system of class i at epoch t. Then the total warranted service to class 1 customers is given by $2n_1/(n_1 + n_2)$ while the granted service is 1. The total discrimination is $\widetilde{D_{(1)}} = 1 - 2n_1/(n_1 + n_2) = (n_2 - n_1)/(n_1 + n_2)$.

Let $p(n_1, n_2)$ be the probability that at an arbitrary epoch there are n_1, n_2 customers in the system. Then the above leads to:

$$\mathbb{E}\{\widetilde{D_{(1)}}\} = \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty} p(n_1, n_2)\frac{n_2 - n_1}{n_1 + n_2}.$$

The expected discrimination for a customer of class 1, $\mathbb{E}\{D_{(1)}\}$ can be derived from (4). Further, note that in the case of server dedication N_1 is independent of N_2 and thus $p(n_1, n_2) = p_1(n_1)p_2(n_2)$ where $p_i(n_i)$ is the probability that $N_i = n_i$. These lead to (and a similar expression of class 2):

$$\mathbb{E}\{D_{(1)}\} = \frac{1}{\lambda_1}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty} p_1(n_1)p_2(n_2)\frac{n_2 - n_1}{n_1 + n_2}; \quad \mathbb{E}\{D_{(2)}\} = \frac{1}{\lambda_2}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty} p_1(n_1)p_2(n_2)\frac{n_1 - n_2}{n_1 + n_2}.$$

Now the difference between these values is :

$$\mathbb{E}\{D_{(2)}\} - \mathbb{E}\{D_{(1)}\} \geq \frac{1}{\lambda_1}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty} p_1(n_1)p_2(n_2)\frac{2(n_1 - n_2)}{n_1 + n_2}.$$

where we used $\lambda_1 \geq \lambda_2$ (which holds for both cases i) and ii)). Note that when $n_1 = n_2$ the term inside the sum is zero. We can therefore sum just for $n_1 \neq n_2$ in the following way:

$$\begin{aligned}
&\mathbb{E}\{D_{(2)}\} - \mathbb{E}\{D_{(1)}\} \\
&\quad \geq \frac{1}{\lambda_1}\left(\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{n_1-1} p_1(n_1)p_2(n_2)\frac{2(n_1 - n_2)}{n_1 + n_2} + \sum_{n_2=1}^{\infty}\sum_{n_1=1}^{n_2-1} p_1(n_1)p_2(n_2)\frac{2(n_1 - n_2)}{n_1 + n_2}\right) \\
&\quad = \frac{1}{\lambda_1}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{n_1-1}\big(p_1(n_1)p_2(n_2) - p_1(n_2)p_2(n_1)\big)\frac{2(n_1 - n_2)}{n_1 + n_2}.
\end{aligned}$$

We now require that $\mathbb{E}\{D_{(2)}\} - \mathbb{E}\{D_{(1)}\} > 0$. Since $n_1 > n_2$, a sufficient requirement is that

$$\frac{p_1(n_1)}{p_1(n_2)} > \frac{p_2(n_1)}{p_2(n_2)} \tag{16}$$

for any $n_1 > n_2 \geq 1$. We now move on to show when this condition holds.

For the $GI/GI/1$ model with LCFS-PR, where interarrival and service requirements are distributed as A and S respectively, it is known that the steady state probability of having

n customers in the system at arbitrary times is geometric, given by $p(k) = \rho(1-\sigma)\sigma^{k-1}, k = 1, 2, \ldots$ where $\rho = \mathbb{E}\{S\}/\mathbb{E}\{A\} = \lambda\mathbb{E}\{S\} < 1, \sigma = (\mathbb{E}\{B\} - 1)/\mathbb{E}\{B\}$, and B is the steady state number of customers served in one busy period (see Núñez-Queija [18] for a review of the literature in this subject).

In the $GI/M/1$ case with FCFS, where S is exponentially distributed with mean $1/\mu$, the same applies, and we have $\mathbb{E}\{B\} = (1 - A^*(\mu))/(1 - 2A^*(\mu))$ where $A^*(s), s \geq 0$ is the Laplace transform of A. This is immediately obtained when noticing that B is 1 with probability $1 - A^*(\mu)$ and is distributed as $B_1 + B_2$, where B_1 and B_2 are i.i.d as B, with probability $A^*(\mu)$. Therefore $\sigma = A^*(\mu)/(1 - A^*(\mu))$.

Using the geometric forms of $p_1(n)$ and $p_2(n)$ we get that (16) is true iff $\sigma_1 > \sigma_2$ which is true iff $A_1^*(\mu_1) > A_2^*(\mu_2)$ where $A_i^*(s), s \geq 0$ is the Laplace transform of $A_i, i = 1, 2$. Since $A_i^*(s)$ is monotone none decreasing this holds when either (i) or (ii) holds. □

Remark 4.1 (Some Comments on Theorem 4.1). 1) In the $M/M/1$ and $D/M/1$ case the conditions (i) and (ii) take the form $\rho_1 > \rho_2$. 2) The conditions (i) or (ii) are sufficient but not necessary. In fact $A_1^*(\mu_1) > A_2^*(\mu_2)$ combined with $\lambda_1 \geq \lambda_2$ is also satisfactory. 3) The final part of the proof (proving that (16) holds when either (i) or (ii) holds) can be achieved in several other ways, e.g. utilizing the fact that $\sigma = A^*(\mu - \mu\sigma)$ and that in our case $A_1^*(s) < A_2^*(s)$. However, we find that the proof above is the more elegant one, and requires the least limitations.

Conjecture 4.1. *Consider a dual class system where each class behaves like a $M/GI/1$ system. Then, if either (i) $\lambda_1 > \lambda_2$ and $S_1 \succeq S_2$, or (ii) $\lambda_1 \geq \lambda_2$ and $S_1 \succ S_2$ then $\mathbb{E}\{D_{(1)}\} < \mathbb{E}\{D_{(2)}\}$.*

We base this conjecture on the fact that the steady state occupancy probabilities in an $M/GI/1$ type system, $p(n)$ can be expressed using the following recursion (see Neuts [17]):

$$p(0) = 1 - \rho$$
$$p(k+1) = \frac{1}{a_0}[\alpha_{k+1}p(0) + \sum_{v=1}^{k} \alpha_{k-v+2}p(v)],$$

where $a = \{a_j\}_0^\infty$ is the probability function of the number of arrivals during a customer's service time and $\alpha_j = \sum_{k=i}^{\infty} a_k$. Let $a^{(i)} = \{a_j^{(i)}\}_0^\infty, i = 1, 2$ denote the probability function for class i, and similarity define $\alpha_j^{(i)}$. Then obviously both (i) and (ii) imply that $a_i^{(1)} \geq a_i^{(2)}, i \geq 1$ and $a_0^{(1)} \leq a_0^{(2)}$, implying $\alpha_i^{(1)} \geq \alpha_i^{(2)}, i \geq 1$ and hinting that $p_1(n_1)/p_1(n_2) > p_2(n_1)/p_2(n_2)$.

We next show that if one just demands that $\rho_1 > \rho_2$ and considers an arbitrary $G/G/1$ system, then the claim of Theorem 4.1 does not necessarily hold.

Theorem 4.2. *In a system like in Theorem 4.1 where each class behaves as a $G/G/1$ system, and: i) The service times are identically distributed for the two classes and ii) $\lambda_1 \geq \lambda_2$, then the claim $\mathbb{E}\{D_{(1)}\} < \mathbb{E}\{D_{(2)}\}$ does not necessarily hold.*

Proof. Consider a system where the service times of both classes are deterministic, equalling one unit. The arrivals of class 1 are deterministic at intervals of one unit (D/D/1) while arrivals to class 2 occur in bulks of size k at inter-arrival time of $m > k$ units. The momentary discrimination of class 1 is given by $\mathbb{E}\{\widetilde{D_{(1)}}\} = \frac{1}{m}\sum_{i=1}^{k}(1 - \frac{2}{i+1})$ and that of class 2 is given by $\mathbb{E}\{\widetilde{D_{(2)}}\} = \frac{1}{m}\sum_{i=1}^{k}(1 - \frac{2i}{i+1})$. It is easy to see that $\mathbb{E}\{\widetilde{D_{(1)}}\} > 0$ and $\mathbb{E}\{\widetilde{D_{(2)}}\} < 0$ and thus $\mathbb{E}\{D_{(1)}\} > 0 > \mathbb{E}\{D_{(2)}\}$. □

4.2 Analysis of Class Discrimination in systems with many Classes

Consider a system with U classes, indexed $1, \ldots, U$, each directed to a dedicated server with a single queue and served according to FCFS. We assume that the arrival process and service times of class i are independent of that of class j ($1 \le i \ne j \le U$). Thus, the steady state occupancy (number in system) of class i is independent of that of class j.

Let $p^{(i)}(n)$ denote the probability that the number of customers of class i present in the system is n. Since class i forms an independent queue, the values of $p^{(i)}(n)$ can be derived from the literature for a wide class of systems. For example, for an $M/M/1$ type queue $p^{(i)}(n) = (1-\rho_i)\rho_i^n$. For an $M/G/1$ queue one can take the Pollaczek-Khinchin Formula of the Laplace-Stieltjes Transform (LST) of the queue occupancy and use standard numerical procedures to derive from it the values of $p^{(i)}(n)$. We will therefore assume that these values are given and show how to derive from them the class discriminations.

Below we demonstrate how to compute the discrimination experienced by class u. Let $p_{(l)}^{(1,2,\ldots,k)}(n)$ denote the steady state probability that the systems of classes $1, 2, \ldots, k$ contain together n customers and l of their servers are busy. Obviously, one should consider only $0 \le l \le k$ and $n \ge l$. Let $p_{(l)}^{(i)}(n)$ denote the same probability for a system consisting of class i only. Using these two probability vectors we can compute $p_{(l)}^{(1,2,\ldots,k,k+1)}()$ from $p_{(l)}^{(1,2,\ldots,k)}()$ and $p_{(l)}^{(k+1)}()$ as follows:

$$p_{(l)}^{(1,2,\ldots,k+1)}(n) = p_{(l)}^{(1,2,\ldots,k)}(n)p_{(0)}^{(k+1)}(0) + \sum_{i=1}^{n} p_{(l-1)}^{(1,2,\ldots,k)}(n-i)p_{(1)}^{(k+1)}(i) \qquad k \ge l > 1 \quad (17a)$$

$$p_{(1)}^{(1,2,\ldots,k+1)}(n) = p_{(1)}^{(1,2,\ldots,k)}(n)p_{(0)}^{(k+1)}(0) + p_{(0)}^{(1,2,\ldots,k)}(0)p_{(1)}^{(k+1)}(n) \qquad l = 1, \quad (17b)$$

$$p_{(0)}^{(1,2,\ldots,k+1)}(0) = p_{(0)}^{(1,2,\ldots,k)}(0)p_{(0)}^{(k+1)}(0) = \prod_{i=1}^{k+1}(1-\rho_i) \qquad l = 0. \quad (17c)$$

Note that the actual convolution is performed in Equation (17a), and we denote this convolution in vector term as $p_{(l-1)}^{(1,2,\ldots,k)}() * p_{(1)}^{(k+1)}()$. Let N be the number of probability elements one keeps for each vector. Then the computational complexity of performing this convolution is $O(N^2)$. Since $1 < l \le k$, exactly $k-1$ such convolutions are required and the overall complexity is $O(kN^2)$.

Applying the above procedure in a recursive mode can yield $p_{(l)}^{(1,2,\ldots,k)}()$ from $p_{(l)}^{(i)}()$, $l = 0, 1$, $i = 1, \ldots, k$ leading to an overall complexity of $O(k^2N^2)$.

Now, the expected momentary discrimination rate for class U can be computed from the vectors $p_{(l)}^{(1,2,\ldots,U-1)}()$ and $p_{(l)}^{(U)}()$ as follows:

$$\mathbb{E}\{\widetilde{D_{(U)}}\} = 0 \cdot P_{(0)}^{(U)}(0) + \sum_{i=1}^{N} p_{(1)}^{(U)}(i) \sum_{l=0}^{U-1} \sum_{j=l}^{N} p_{(l)}^{(1,2,\ldots,U-1)}(j) \left(1 - (l+1)\frac{i}{i+n}\right). \qquad (18)$$

The computational complexity of this expression, plus the complexity of recursively computing $p_{(l)}^{(1,2,\ldots,U-1)}()$ is $O(U^2N^2)$.

If one wishes to compute the expected value of the momentary discrimination rate for all U classes the computational complexity is $O(U^3N^2)$ steps.

Finally, the expected discrimination for each class can be derived from Equation 18 and from the discrimination version of Little's Law (4).

4.3 A Practical Example – The Restroom Queueing Problem

A common classification and prioritization mechanism used in queues of humans is to classify customers to servers by some server-convenient criteria. For example, in customs queues used in airports, often separate servers (and queues) are used for local citizens and for visitors. Perhaps a more common example is the partition of public restrooms by gender. This can be modeled by a system where each gender is assigned a dedicated set of servers and a single queue. A very common situation is where the delay experienced in the women's queue is much larger than that in the mens' queue. The fairness of this commonly used system has not been quantitatively evaluated to date.

The analytic approach presented in this paper can be extended to apply to such single-queue multi-server or multi-queue multi-server systems as well. Such extension is left for future work. However, since our focus in this work is on customer classification we employ the approach to numerically examine an example of the restroom queue practice.

Several questions need to be addressed in this context. 1) How servers should be assigned to each class in this scenario as to lead to maximally fair scheduling. 2) How fair are current practices. 3) How high is the class-discrimination experienced under various server assignments. For the sake of example we consider the typical queueing problem in public restrooms at theaters, where a large number of customers arrives concurrently (beginning of the theater break) at the servers, queues up and gets served at some order. We will assume that the total number of servers is 12 and for simplicity the number of customers of each gender equals $12j$, j even. The service times of class 1 and class 2 are deterministically 2 and 1 time units respectively. We consider two intuitive and common server assignment policies: 1) Proportional assignment where the number of servers is proportional to the service requirements, that is 8 and 4 servers to class 1 and 2, respectively, and 2) Equal assignment, namely 6 servers to each class.

We first analyze the proportional allocation policy and derive the class discrimination of class 2. Since service times are deterministic we can track the system along time slots where the slot size equals the service time of class 2 (short jobs). Assuming that there are $12j$

jobs of each type at time zero then the number of slots is $3j$ and the number of short jobs present, counting from the last slot backwards, is given by $4i$, $i = 1, ..., 3j$, and the number of long jobs is $8, 8, 16, 16, 24, ..., 12j, 12j$. Thus the overall warranted service of class 2 (along the $3j$ slots) is given by $12 \sum_{i=1}^{3j/2} \frac{2i}{4i} + \frac{2i-1}{4i-1}$. Recalling that the total granted service to class 2 is $12j$, and that there are $12j$ customers, the class discrimination of class 2 is given by

$$\mathbb{E}\{D_{(2)}\} = 1 - \frac{3}{4} - \frac{1}{j}\sum_{i=1}^{3j/2} \frac{2i-1}{4i-1} \approx \frac{1}{4} - \frac{1}{j}\int_{x=1}^{3j/2} \frac{2x-1}{4x-1}dx] = \frac{1}{4} - \frac{6j + \ln 3 + \ln(6j-1) - 4}{8j}, \tag{19}$$

which for large values of j tends to $-1/2$, while $\mathbb{E}\{D_{(1)}\} = -\mathbb{E}\{D_{(2)}\} = 1/2$.

Next we analyze the system under equal server allocation. Similar analysis yields the overall warranted service of class 2 to be: $6 \sum_{i=1}^{j} \frac{2i-1}{j+3i-1} + \frac{2i}{3i+j}$, which yields the class discrimination:

$$\mathbb{E}\{D_{(1)}\} \approx 1 - \frac{1}{9j}\left(-12 + 12j - 2j\ln 4 + 2j \ln \frac{3+j}{j} + (1+2j)\left(\ln(2+j) - \ln(4j-1)\right)\right) \tag{20}$$

which for large values of j tends to $(\log 256 - 3)/9 \approx 0.28$.

The analysis therefore reveals that under our fairness model, neither proportional assignment nor equal assignment is fair. Under proportional assignment the short jobs are negatively discriminated due to the relatively small number of servers allocated to them (out of proportion to their part in the population). Under equal assignment the long jobs are negatively discriminated due to the considerable amount of time during which they form a large majority of the presence (most short jobs are gone earlier) while receiving only half of the resources. The values of discrimination under both allocations are considerably high. The "optimal" point of operation is therefore at the 7:5 allocation in which a simulation shows that $\mathbb{E}\{D_{(2)}\} \approx 0.06$.

5 Concluding Remarks

A common practice in queueing systems is to classify customers into classes and grant the service based on these classes. Fair treatment of customers in such systems is highly important. Our study dealt with the issues of fairness and class discrimination, where we focused on the practices of class prioritization and resource dedication to classes. To address class discrimination we introduced a new metrics called class-discrimination, and used it in addition to the RAQFM measure for system unfairness.

We established several general results for these systems, such as: 1) The (weighted) value of class discrimination is always bounded by the system unfairness; that is, a class cannot be highly discriminated if the overall system unfairness is low. 2) If service order is not based on service times, short jobs are negatively discriminated. 3) In a preemptive priority system, the highest priority class always enjoys positive discrimination. 4) In a one-server per-class

system of the GI/M/1 type, a class whose service times are larger and arrival intervals shorter is guaranteed to benefit positive discrimination.

The results derived in this work can serve for two purposes. First, the simpler results, which might sound "intuitive" to many, should be used to build confidence both in the RAQFM unfairness metrics and the class discrimination metrics; both are very new to the queueing theory world and require examination and trust building. Second, once the confidence is built, the measures can be used to evaluate and study systems where the results are not that clear. For example, we used them to evaluate the "restroom queue problem" and demonstrated that neither of the "intuitive" server assignment rules, namely equal assignment and proportional assignment, minimize class discrimination. We thus expect that both the RAQFM unfairness measure and the class discrimination measure should be useful in evaluating the fairness aspects of a variety of systems that are used in practice.

Lastly, the study of queue fairness is yet in its infancy and many subjects, including fairness in many server cases, fairness in a queueing network and others, remain untouched. Such subjects should be treated in future research.

References

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, third edition, 1999.

[2] B. Avi-Itzhak. Preemptive repeat priority queues as a special case of the multipurpose server problem – I. *Oper. Res.*, 11(4):597–609, 1963.

[3] B. Avi-Itzhak. Preemptive repeat priority queues as a special case of the multipurpose server problem – II. *Oper. Res.*, 11(4):610–619, 1963.

[4] B. Avi-Itzhak and H. Levy. On measuring fairness in queues. *Advances in Applied Probability*, 36(3):919–936, September 2004.

[5] B. Avi-Itzhak, H. Levy, and D. Raz. Quantifying fairness in queueing systems: Principles, approaches and applicability. Technical Report RRR-25-2005, RUTCOR, Rutgers University, August 2005. URL `http://rutcor.rutgers.edu/pub/rrr/reports2005/25_2005.pdf`. Submitted.

[6] B. Avi-Itzhak and P. Naor. On a problem of preemptive priority queuing. *Oper. Res.*, 9(5):664–672, 1961.

[7] E. Brosh, H. Levy, and B. Avi-Itzhak. The effect of service time variability on job scheduling fairness. Technical Report RRR-24-2005, RUTCOR, Rutgers University, July 2005. URL `http://rutcor.rutgers.edu/pub/rrr/reports2005/24.pdf`.

[8] D. R. Cox and W. L. Smith. *Queues*. Methuen/Wiley, London, 1961.

[9] T. A. Davis. *UMFPACK Version 4.0 User Guide*. Dept. of Computer and Information Science and Engineering, Univ. of Florida, Gainesville, Florida, 2002.

[10] A. van Harten and A. Sleptchenko. On multi-class multi-server queueing and spare parts management. Technical Report WP-49, BETA publication, University of Twente, Enschede, The Netherlands, 2000.

[11] A. van Harten, A. Sleptchenko, and M. C. van der Heijden. On multi-class multi-server queue with preemptive priorities. Technical Report WP-77, BETA publication, University of Twente, Enschede, The Netherlands, 2003.

[12] N. K. Jaiswal. *Priority Queues*. Academic Press, New York, 1968.

[13] O. Kella and U. Yechiali. Waiting times in the non-preemptive priority M/M/c queue. *Commun. Statist.-Stochastic Models*, 1(2):257–262, 1985.

[14] O. Kella and U. Yechiali. Priorities in M/G/1 queue with server vacations. *Naval Research Logistics*, 35:23–24, 1988.

[15] L. Kleinrock. *Queueing Systems, Volume 2: Computer Applications*. Wiley, 1976.

[16] J. D. C. Little. A proof of the queueing formula $l = \lambda w$. *Operations Research*, 9:380–387, 1961.

[17] M. F. Neuts. Algorithms for the waiting time distributions under various queue disciplines in the M/G/1 queue with service time distributions of phase type. In M. F. Neuts, editor, *Algorithmic Methods in Probability, TIMS studies in the Managment Sciences*, volume 7, pages 177–197. North-Holland Publishing Co., London, 1977.

[18] R. Núñez-Queija. Note on the GI/GI/1 queue with LCFS-PR observed at arbitrary yimes. *Probability in the Engineering and Informational Sciences*, 15:179–187, 2001.

[19] A. Rafaeli, G. Barron, and K. Haber. The effects of queue structure on attitudes. *Journal of Service Research*, 5(2):125–139, 2002.

[20] A. Rafaeli, E. Kedmi, D. Vashdi, and G. Barron. Queues and fairness: A multiple study investigation. Technical report, Faculty of Industrial Engineering and Management, Technion. Haifa, Israel. Under review, 2003. URL `http://iew3.technion.ac.il/Home/Users/anatr/JAP-Fairness-Submission.pdf`.

[21] D. Raz, B. Avi-Itzhak, and H. Levy. Classes, priorities and fairness in queueing systems. Technical Report RRR-21-2004, RUTCOR, Rutgers University, June 2004. URL `http://rutcor.rutgers.edu/pub/rrr/reports2004/21_2004.pdf`. Submitted.

[22] D. Raz, B. Avi-Itzhak, and H. Levy. Fair operation of multi-server and multi-queue systems. Submitted for publication, 2004. URL `http://www.cs.tau.ac.il/~davidraz/mult-d9a.pdf`.

[23] D. Raz, H. Levy, and B. Avi-Itzhak. A resource-allocation queueing fairness measure. In *Proceedings of Sigmetrics 2004/Performance 2004 Joint Conference on Measurement and Modeling of Computer Systems*, pages 130–141, New York, NY, June 2004. (*Performance Evaluation Review*, 32(1):130-141).

[24] W. Sandmann. A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement. Accepted for the 1st Euro NGI Conference on Next Generation Internet Networks Traffic Engineering, April 2005.

[25] E. Sherry-Gordon. *New Problems in Queues: Social Injustice and Server Production Management*. PhD thesis, MIT, May 1987.

[26] A. Sleptchenko. *Integral Inventory Control in Spare Parts Networks with Capacity Restrictions*. PhD thesis, University of Twente, 2002.

[27] H. Takagi. *Queueing Analysis, A Foundation of Performance Evaluation Volume 1: Vacation and Priority Systems (Part 1)*. North-Holland, Amsterdam, The Netherlands, 1991.

[28] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM Sigmetrics 2003 Conference on Measurement and Modeling of Computer Systems*, pages 238 – 249, San Diego, CA, June 2003.