Class Treatment in Queueing Systems: Discrimination and Fairness Aspects

David Raz*
Holon Institute of Technology
Holon, Israel
davidra@hit.ac.il

Hanoch Levy†
Computer Engineering and Networks Lab
ETH, Zurich, Switzerland
hanoch@tik.ee.ethz.ch

Benjamin Avi-Itzhak
RUTCOR, Rutgers University
New Brunswick, NJ, USA
aviitzha@rutcor.rutgers.edu

ABSTRACT

Customer classification and prioritization are commonly utilized in applications to provide queue preferential service. Their fairness aspects, which are inherent to any preferential system and highly important to customers, have not been fully studied and quantified to date. We use the recently proposed Resource Allocation Queueing Fairness Measure (RAQFM), and a newly introduced metric called class discrimination, which is based on RAQFM, to analyze such systems and derive their relative fairness values as well as the discrimination experienced by the various classes. Specifically, we study two practices, commonly used in public facilities as well as in computer systems: *class prioritization* and *dedication of resources to classes*.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Performance Attributes—*Fairness*; F.2.2 [**Nonnumerical Algorithms and Problems**]: Sequencing and Scheduling; G.3 [**Probability and Statistics**]: Queuing Theory

General Terms

Performance, Measurement

Keywords

Fairness, Discrimination, Prioritization, Multiple Classes, Job Scheduling, Resource Allocation, Unfairness

1. INTRODUCTION

*A large part of this work was done while Raz was with the School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel

†On leave of absence from Tel-Aviv University.

ValueTools 2008 October 21–23, 2008, Athens, Greece

1.1 Forward and Motivation

Customer classification and prioritization mechanisms are commonly 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". As shown in recent Experimental Psychology studies [17, 18] fairness in the queue is very important to people, perhaps not less than the wait itself.

Despite this fact, the general fairness aspects of prioritization and classification in queueing systems and their quantitative evaluation have been studied only little, especially compared to the volume of literature on the utility of these systems. The effect of these mechanisms on fairness cannot be quantitatively accounted for in the queueing analysis of many daily life applications.

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 in [22]. The measure is based on the application of the basic Rawlsian social and economical justice conception ([19]), that equally needy members of a group should share equally the resources available to the group. 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, either positive or negative. An advantage of RAQFM is that it is analyzable, as well as reactive to both seniority and size differences among jobs. 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 lower 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.

The importance of this work is that it is perhaps the first attempt to evaluate the fairness aspects of these two mechanisms and conduct it in quantitative manner. The results

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

derived and the techniques developed can be used for conducting fairness evaluation of such practical systems in a variety of applications.

1.2 Overview and Main Contributions

In analyzing the fairness aspects of prioritization and classification mechanisms we will first be interested in the overall fairness of the system and how the queueing mechanism affect it. Nonetheless, dealing with classes we realize that an additional important quantity is the relative treatment given to the different classes. We thus introduce a new metric, called *class discrimination*, which is a variant of RAQFM, and which accounts for the expected discrimination experienced by the customers of a certain class. Analysis of this metric, its properties and its relations to the RAQFM unfairness measure is given in Sections 2.3 and 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. There is a large body of literature on such priority schemes (e.g. textbooks [10, 13, 26]) where the focus in evaluating system performance is on the system expected waiting time, or the mean waiting cost, under linear cost parameters. Optimization of the system with non-preemptive priorities, based on this performance objective, shows (e.g. [5, 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 with providing a "justification" for short job prioritization. Of course there is little need to justify such prioritization in terms of performance, as such justification was provided as early as 1996 in the case of SRPT, see [25]. Our justification, on the other hand, is in terms of class discrimination. We show that for any service policy that selects customers for service *independently* of their required *service times*, the discrimination experienced by a customer is monotone non-decreasing with respect to its service time. This means that an implicit discrimination is applied in favor of the long jobs. This implies 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* (*lowest*) priority class is always positive (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) provide numerical results from an analysis of unfairness for a system with two customer classes and class prioritization. The results show that in many cases, though not in all, prioritization of the short jobs over the long jobs leads to higher fairness than that of First Come First Server (FCFS).

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 set of servers is dedicated to each class. Examples of such systems include call centers with multiple classes of customers, and airport passport control systems. 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. We first show that for a system consisting of two $\mathsf{GI/M/1}$ queues, if either i) The inter-arrival time of class 1 is stochastically larger 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: We show a counetr example for $\mathsf{G/G/1}$ queues. Second, we deal with how to compute class discrimination in a system with several classes, either several $\mathsf{M/M/1}$'s or several $\mathsf{M/GI/1}$'s. We propose an algorithm that computes class discrimination in a computation complexity which is polynomial in the number of classes. We conclude with some numerical results for stochastic service times and for a special "rush hour scenario". The results show that sometimes neither allocating the resources equally, nor allocating them proportionally to the service times lead to the least unfairness.

Concluding remarks are given in Section 5.

1.3 Related Work

We now briefly survey some related works on the subject of fairness. For lack of space we do not go into details. A broader survey is available in [20].

Fairness in our work corresponds to fairness to specific jobs in the system, and is denoted *job fairness*. A comprehensive overview of alternative measures for job fairness is available in two recent surveys [3, 27]. Example approaches include Skips and Slips [7], Expected Slowdown [28], Slowdown Fairness [2], and Discrimination Frequency [24]. None of the studies that proposed these approaches dealt explicitly with prioritization and classification mechanisms. It is an open question whether these alternative approaches can be applied to derive and study the relative fairness of the mechanisms studied in this work. This is the subject of current work.

Fairness has been excessively treated in contexts different than that of job fairness. In the area of flow control the best known notion in this area is that of *Max-Min Fairness* (Starting with [9] and used by many afterwards), followed by *Proportional Fairness* ([11] and others). A large volume of literature also exists on *Weighted Fair Queueing* (e.g. [6], [8]). Another area where related works are starting to appear is parallel job schedulers, e.g. [23].

2. SYSTEM MODEL AND MEASURES

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 s_l denote the service requirement of C_l, measured in time units. Each customer belongs to one of U classes, indexed $1, 2, \ldots, U$. Let $v(C_l)$ be the class to which C_l belongs. The arrival rate and expected service requirement of class u customers are are λ_u and $1/\mu_u$ respectively, where $\sum_{u=1}^{U} \lambda_u = \lambda$. An order of priorities is assigned to the classes, where lower class index means higher priority.

In our discussion, when we mention the *Preemptive Priority* class of scheduling policies, the order of service within each class of customers is FCFS, and preempted customers

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

return to the head of the queue of their class. For discussion of this, and other variants, see [26, Sec. 3.4].

2.2 System Unfairness Measure: RAQFM

RAQFM was proposed in [22] and defined there for a single server with fixed rate. Below we slightly generalize it to multiple servers with time varying rate, a generalization which was also used in [21]. RAQFM evaluates the unfairness in the system by first providing a metric of the discrimination of a customer, and then using a summary formula to evaluate the system unfairness. The discrimination is evaluated as follows: The fundamental assumption is that at each epoch, all customers present in the system deserve an *equal share* of the *total service available*. If we let $0 \leq \omega(t) \leq M$ denote the total service rate available at epoch t, then the fair share, called the instantaneous *warranted service* rate, is $\omega(t)/N(t)$, where $N(t)$ is the number of customers in the system at epoch t.

The choice of $\omega(t)$ depends on the specific application, and on what is perceived as the service rate available to customers. One simple option is to use $\omega(t) = M$, the number of servers. However, this option is less appropriate in many situations, for example, a multiple server system where only one job is present, and it is physically impossible to serve that job with more than one server. We therefore choose to focus on a second, more realistic option; the total service rate available at epoch t is the total service rate physically granted to customers present at the system at that epoch. Other choices of $\omega(t)$ are also possible, but as these issues are more pronounced in multi-server multi-queue systems we discuss them in depth in a study that focuses on these systems ([21]).

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

The instantaneous 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

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

For 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)

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

Let D be a the discrimination experienced by an arbitrary customer C (a random variable). An important property of RAQFM (shown for a non-idling server in [22], 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 instantaneous discrimination rate that when taken over all customers, sums to zero at any time epoch t. At idling epochs both $\omega(t)$ and $\sigma_l(t)$ are zero. □

Intuitively, this theorem means that RAQFM has the "zero sum" property, where positively discriminating a customer must be done on the account of negatively discriminating other customers, in equal total amount.

Since $\mathbb{E}\{D\} = 0$, and following [22], the summary formula used to evaluate the unfairness of the system is the variance of the discrimination, which is equal to the second moment of the discrimination, namely $\mathbb{E}\{D^2\}$. Other possible approaches are to use the expected absolute value $\mathbb{E}\{|D|\}$ or to use higher moments and cumulants (see for example [29]).

2.3 Class Discrimination and its Basic Properties

For systems with customer classification it is important to evaluate the comparable, or relative treatment given to each class. To this end we introduce the notion of *class discrimination* which we define here on the basis of RAQFM, and relates to the discrimination experienced by a certain class of the population. For a class u the discrimination D experienced by an arbitrary customer C, when the system is in steady state, is a random variable denoted $D_{(u)} = D|\upsilon(C) = u$. Our interest will be in the expected discrimination experienced by u's customers, namely $\mathbb{E}\{D_{(u)}\}$, termed *Class Discrimination*.

A second useful notion is that of *class discrimination rate*. The *instantaneous 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 $\tilde{D}_{(u)}(t) = \sum_{\upsilon(C_l)=u} c_l(t)$ denote this variable and let $\tilde{D}_{(u)} = \lim_{t\to\infty} \tilde{D}_{(u)}(t)$ be a random variable denoting the instantaneous discrimination rate of class u when the system is in steady state. Taking expectation of this variable we get the class discrimination rate $\mathbb{E}\{\tilde{D}_{(u)}\}$.

We observe that the relationship between the variables $D_{(u)}$ and $\tilde{D}_{(u)}$ is analogous to the equilibrium relationship between the variables *customer delay* (delay experienced by an arbitrary customer) and *number of customers in the system* (number of customers present at an arbitrary moment) in a stationary queueing system. While the former is more appropriate to describe the customer's perception, the latter might be more appropriate to describe the system's state. We therefore choose to focus on the former.

Using Brumelle's theorem $H = \lambda G$, where H and G are respectively time and customer averages of the same quantity ([4]), we derive

$$\mathbb{E}\{\tilde{D}_{(u)}\} = \lambda_u \mathbb{E}\{D_{(u)}\}. \tag{4}$$

Note that while class discrimination was defined for a system in steady state it can also be computed for a given scenario, where instead of using the expected value of a random variable one uses the statistical average of a given realization. For lack of space we do not give the full definition.

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 class discrimination. We first show

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

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

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

□

Note a similar result in the area of bandwidth sharing [12]. Our result can be viewed as an extension of this work.

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.

Note that the opposite is not correct. If class discrimination of each class u obeys $\mathbb{E}\{D_{(u)}\} = 0$, then the system unfairness, $\mathbb{E}\{D^2\}$ can still be positive.

For 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 class discrimination is zero for both classes.

2.4.1 Practical Implications

The practical implications of these results are: 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, while 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 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, and that while prioritization can guarantee positive discrimination to the highest priority class and negative discrimination to the lowest priority class, it cannot guarantee monotonicity in discrimination. Lastly, we provide numerical results from evaluating the unfairness in single server systems with preemptive priority.

3.1 Prioritizing Short Jobs is Justified

DEFINITION 3.1 (STOCHASTIC DOMINANCE). *Consider nonnegative random variables X_1, X_2 whose distribution functions 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 C_l be a customer, and let $D_l^{(s_l)}$ be a random variable denoting the discrimination in steady state of C_l as function of its service time s_l. Then $D_l^{(s_l)}$ is monotone non-decreasing with respect to 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. □

This simple theorem, stated in terms of deterministic service requirements, can also be stated using stochastic service requirements, i.e. if the customer's service requirements are stochastic variables S_l and S_l', and $S_l' \succ S_l$ then $D^{(S_l)} \succ D^{(S_l')}$ and clearly $\mathbb{E}\{D^{(S_l')}\} \geq \mathbb{E}\{D^{(S_l)}\}$.

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

In conclusion, we have shown that service policies that do not give preferential 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.

THEOREM 3.2. *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 $\mathbb{E}\{D_{(U)}\} \leq 0$.*

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

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

$$\tilde{D}_{(1)}(t) = \begin{cases} N_1(t) - \frac{\omega(t)N_1(t)}{N(t)} & N_1(t) \leq M \\ M - \frac{MN_1(t)}{N(t)} & N_1(t) > M \end{cases} = \begin{cases} N_1(t)\left(1 - \frac{\omega(t)}{N(t)}\right) & N_1(t) \leq M \\ M\left(1 - \frac{N_1(t)}{N(t)}\right) & N_1(t) > M \end{cases},$$

which is nonnegative since $\omega(t) \leq N(t)$ and $N_1(t) \leq N(t)$.

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

As for $\tilde{D}_{(U)}(t)$, 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, $\tilde{D}_{(U)}(t) = -N_U(t)M/N(t)$ 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

$$\tilde{D}_{(U)}(t) = \omega_U(t) - \frac{N_U(t)M}{N(t)} = \frac{\omega_U(t)N(t) - N_U(t)M}{N(t)}. \quad (7)$$

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 (7) yields

$$\tilde{D}_{(U)}(t) = \frac{\omega_U(t)(M + N'(t)) - (\omega_U(t) + N'(t))M}{N(t)} = \frac{(\omega_U(t) - M)N'(t)}{N(t)} < 0,$$

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

The important thing about Theorem 3.2 is that the the most prioritized class has nonnegative discrimination, even if the customers are extremely small. This means that at least for the first priority class, certain discrimination can be guaranteed.

Having shown that the discrimination of the most prioritized class is always non-negative, and that of the least prioritized class is always non-positive, one might expect that the discrimination is monotonic with respect to 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 discrimination for the four classes. The results were achieved through simulation, although similar results can be achieved through numerical analysis, using the method presented in [20].

Observe that when the service requirement of class 2 is equal to that of the other classes, the class discrimination is monotonic with respect to 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.

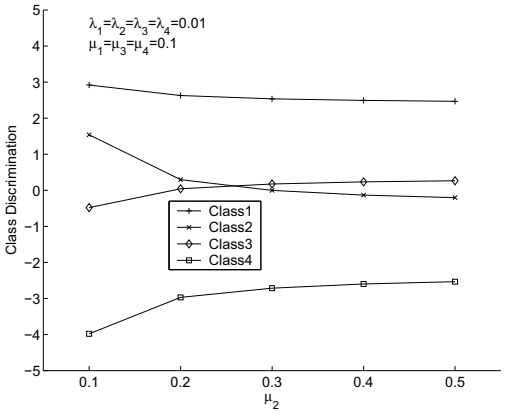


Figure 1: The Effect of Preemptive Priority on Four Classes of Customers. $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.01, \rho_1 = \rho_3 = \rho_4 = 0.1$.

One interesting case is where there are two classes, and the prioritized class has infinitesimal service requirements (i.e. $1/\mu_1 \to 0$). It is easy to see that in this case $\mathbb{E}\{D_{(1)}\} \to 0$. Interestingly if $\lambda_1/\lambda_2 \nrightarrow \infty$ it is also easy to see that $\mathbb{E}\{D_{(2)}\} \to 0$. This can be seen immediately if one considers the following conservation law, which is consequence of Brumelle's theorem ([4]):

$$\frac{\sum_{u=1}^{U} \lambda_i \mathbb{E}\{D_{(i)}\}}{\sum_{u=1}^{U} \lambda_i} = 0. \quad (8)$$

It is interesting to draw the similarity of this law to the conservation law regarding the waiting time of customer classes in a single server systems (see [13, Chap. 3.4]).

3.3 Fairness Single Server Systems with Preemptive Priority

For lack of space we do not bring the full analysis, which can be found in [20]. The analysis is based on the methodology developed in [22] which is extended to deal with classes and prioritization. It is immediately applicable to systems with exponential service and inter-arrival times, and can be generalized to arbitrary phase-type distributions.

We go directly to presenting numerical results. One specific case of interest is the following: suppose that a call center services two types of customers, one requiring only a brief approval, and one requiring the full attention of a service person for several minutes. It is common to suggest, due to fairness reasons, that customers with shorter service requirement should be served ahead of other customers (this situation applies also to other fields where customers of different service requirements are served by the same servers). 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 mini-

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

mum 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.

To this end, we examine the unfairness as a function of the service difference between the customer classes, expressed by the mean service time ratio μ_1/μ_2. As demonstrated in [22], the unfairness of the system is sensitive to the utilization ρ. Therefore, we maintain constant utilization $\rho = 0.8$, independently of c. For simplicity, the evaluation is done for equal arrival rates of $\lambda_1 = \lambda_2 = 0.1$.

Figure 2(a) depicts the unfairness of the system, and compares it to the unfairness in a FCFS schedule. For the FCFS system one can observe that since the system does not prefer either of the classes they are interchangeable, and the unfairness is therefore symmetric with respect to μ_1/μ_2.

For the prioritized system one may observe that: 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) In the right side of the figure, where the shorter jobs receive priority, the system unfairness slightly increases with the service requirement ratio, but much less than in the FCFS case. Comparing the systems, we observe that: 3) When class 1 customers have longer expected service requirement it is less fair, system-wise, to give them priority. 4) 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.

Figure 2(b) and Figure 2(c) shows results where the service distribution has the same expected value, and the distribution is Erlang-10 and Coxian-2 respectively (see e.g. [1, Chap. 2])). For the Erlang-10 distribution we have a coefficient of variance of $1/\sqrt{10}$ (The coefficient of variance of a random variable X is $Var\{X\}/\mathbb{E}\{X\}$). For the Coxian-2 distribution we use the settings suggested by [14], to achieve a coefficient of variance of $\sqrt{10}$: to achieve an expected value of S with a coefficient of variance of C use $\mu_1 = 2/S, p_1 = 1/(2C^2), \mu_2 = 1/(C^2 S)$.

One may observe that the behavior is very similar to the behavior observed for Exponential distribution. One difference is that for the priority service, in the right side of the figures, for the Erlang-10 distribution there is almost no increase in system unfairness, and in the Coxian-2 there is a large increase. This agrees with our expectations since for Erlang-10 there is very little variability within the class, and the opposite for Coxian-2.

To conclude this section, we observe that each distribution is characterized by a threshold. If the ratio between the mean non-prioritized job size and the mean prioritized job size is below this threshold, it is more fair to serve the customers in FCFS manner. Otherwise, the priority manner is more fair.

Recall the call center example, 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

In this section we deal with the dedication of resources to classes. We consider systems where each class is assigned a dedicated set of servers and a FCFS queue. We focus on analyzing the class discrimination in these systems. In Section 4.1 we analyze systems with two classes, where each class has a single dedicated server. In Section 4.2 we provide an algorithmic approach for deriving the class discrimination for more general systems. In Section 4.3 we show numerical examples.

4.1 Dominance Results for 2 Class Systems

In this section we analyze systems with two classes, where each class has a single dedicated server. We show that in many cases equal resource allocation results in negatively discriminating the heavier loaded class.

THEOREM 4.1. *Consider a system with two classes, where each class is served by a single server on a single GI/M/1 queue, in a manner that does not take size or remaining size into account. 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 $\tilde{D}_{(u)}$ be a random variable denoting the total instantaneous discrimination rate to class u customers at steady state. $\tilde{D}_{(1)}$ can be derived by conditioning on the system state and examining three cases: 1) Case 1—server 1 is idle: In this case no class 1 customers are present in the system and thus $\tilde{D}_{(1)} = 0$. 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 $\tilde{D}_{(1)} = 0$. 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. 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 $\tilde{D}_{(1)} = 1 - 2n_1/(n_1+n_2) = (n_2 - n_1)/(n_1 + n_2)$.

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:

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

The class discrimination for 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:

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

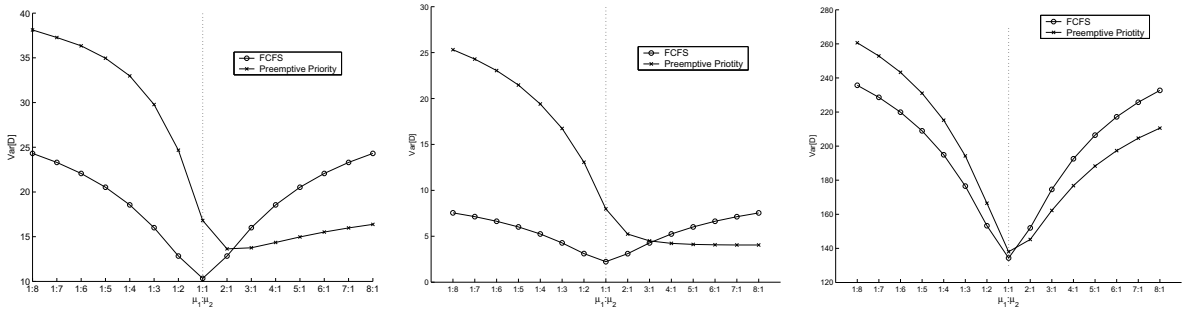
$$\mathbb{E}\{D_{(2)}\} = \frac{1}{\lambda_2}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty} p_1(n_1)p_2(n_2)\frac{n_1 - n_2}{n_1 + n_2}.$$

Now the difference between these values is :

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

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:

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344



(a) Exponential Distribution (b) Erlang-10 Distribution (c) Coxian-2 Distribution

Figure 2: Unfairness in a Single Server System with Two Customer Classes, $\lambda_1 = \lambda_2 = 0.1, \rho = 0.8$

$$\begin{aligned}
&\mathbb{E}\{D_{(2)}\} - \mathbb{E}\{D_{(1)}\} \\
&= (\frac{1}{\lambda_1} + \frac{1}{\lambda_2})\Big(\sum_{n_1=1}^{\infty} \sum_{n_2=1}^{n_1-1} p_1(n_1)p_2(n_2)\frac{n_1 - n_2}{n_1 + n_2} \\
&\quad + \sum_{n_2=1}^{\infty} \sum_{n_1=1}^{n_2-1} p_1(n_1)p_2(n_2)\frac{n_1 - n_2}{n_1 + n_2}\Big) = \\
&\quad (\frac{1}{\lambda_1} + \frac{1}{\lambda_2})\times \\
&\sum_{n_1=1}^{\infty} \sum_{n_2=1}^{n_1-1} \big(p_1(n_1)p_2(n_2) - p_1(n_2)p_2(n_1)\big)\frac{n_1 - n_2}{n_1 + n_2}.
\end{aligned}$$

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

$$\frac{p_1(n_1)}{p_1(n_2)} > \frac{p_2(n_1)}{p_2(n_2)} \tag{9}$$

for any $n_1 > n_2 \geq 1$, and we move on to show when this condition holds.

For the GI/GI/1 model with Preemptive Last Come First Served, 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 [16] for a review of the literature on 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 (9) 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 non-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 cases conditions (i) and (ii) take the form $\rho_1 > \rho_2$. 2) Conditions (i) or (ii) are sufficient but not necessary. In fact $A_1^*(\mu_1) > A_2^*(\mu_2)$ is also satisfactory. 3) The final part of the proof (proving that* (9) *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 more elegant, and requires the least limitations.*

Corollary 4.1. *Under the same conditions as in Theorem 4.1, $\mathbb{E}\{D_{(1)}\} < 0, \mathbb{E}\{D_{(2)}\} > 0$.*

This is clear from Theorem 4.1 and (8).

We conjecture that a similar theorem can also be proved for M/GI/1 systems. This is based 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 [15]):

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

where $a = \{a_j\}_0^\infty$ is the probability function of the number of arrivals during a customer's service time and $\alpha_j = \sum_{k=j}^{\infty} a_k$. Let $a^{(i)} = \{a_j^{(i)}\}_0^\infty, i = 1, 2$ denote the probability function for class i, and similarly 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 $p_1(n_1)/p_1(n_2) > p_2(n_1)/p_2(n_2)$.

The claim of Theorem 4.1 does not necessarily hold if one demands only that $\rho_1 > \rho_2$ and considers an arbitrary G/G/1 system. Consider for example 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) and the arrivals to class 2 occur in bulks of size k at inter-arrival time of $m > k$ units. The instantaneous discrimination of class 1 is given by $\mathbb{E}\{\tilde{D}_{(1)}\} = \frac{1}{m}\sum_{i=1}^{k}(1 - \frac{2}{i+1})$ and that of class 2 is given by $\mathbb{E}\{\tilde{D}_{(2)}\} = \frac{1}{m}\sum_{i=1}^{k}(1 - \frac{2i}{i+1})$. It is easy to see that $\mathbb{E}\{\tilde{D}_{(1)}\} > 0$ and $\mathbb{E}\{\tilde{D}_{(2)}\} < 0$ 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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

according to FCFS. We assume that the arrival process and service times of class i are independent of that of class j $(1 \leq i \neq j \leq U)$. 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 $p^{(1,2,\ldots,k)}(n,l)$ denote the steady state probability that the system of classes $1,2,\ldots,k$ contains together n customers and l of their servers are busy. Obviously, one should consider only $0 \leq l \leq k$ and $n \geq l$. We can now compute $p^{(1,2,\ldots,k,k+1)}()$ from $p^{(1,2,\ldots,k)}()$ and $p^{k+1}()$ as follows:

$$p^{(1,2,\ldots,k+1)}(n,l) = p^{(1,2,\ldots,k)}(n,l)p^{k+1}(0) + \sum_{i=1}^{n} p^{(1,2,\ldots,k)}(n-i,l-1)p^{k+1}(i) \qquad 1 \leq l \leq k \quad (10a)$$

$$p^{(1,2,\ldots,k+1)}(0,0) = p^{(1,2,\ldots,k)}(0,0)p^{k+1}(0) = \prod_{i=1}^{k+1}(1-\rho_i) \qquad l = 0. \quad (10b)$$

Note the convolution in (10a). 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 $1 \leq l \leq k$ the overall complexity for evaluating $p^{(1,2,\ldots,k+1)}()$ from $p^{(1,2,\ldots,k)}()$ is $O(kN^2)$. Applying this procedure recursively for all classes up to $u-1$ leads to an overall complexity of $O(u^2N^2)$.

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

$$\mathbb{E}\{\tilde{D}_{(u)}\} = 0 \cdot P^u(0) + \sum_{i=1}^{N} p^u(i) \sum_{l=0}^{u-1} \sum_{j=l}^{N} p^{(1,2,\ldots,u-1)}(j,l) \left(1 - (l+1)\frac{i}{i+n}\right). \quad (11)$$

The total computational complexity is therefore $O(u^2N^2)$.

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

Finally, the class discrimination can be derived from (11) and (4).

4.3 Numerical Results

Several questions need to be addressed in this context of resource dedication to classes. 1) How servers should be assigned to each class 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.

We start with a simple stochastic service example. Consider a system where the Poisson arrival rates of the two classes are identical $\lambda_1 = \lambda_2 = 0.005$ and the service times are exponential where the mean of class 1 doubles that of class 2: $\mu_1 = 0.12, \mu_2 = 0.24$. We consider a system consisting of 12 servers and evaluate system unfairness and class discrimination as a function of the server assignment policy (dedication of $k : 12-k$ where k is the dedication to class 1 and $12-k$ to class 2). Figure 3 depicts these results. Note that class discriminations are plotted against the left y-axis, while unfairness is plotted against the right y-axis.

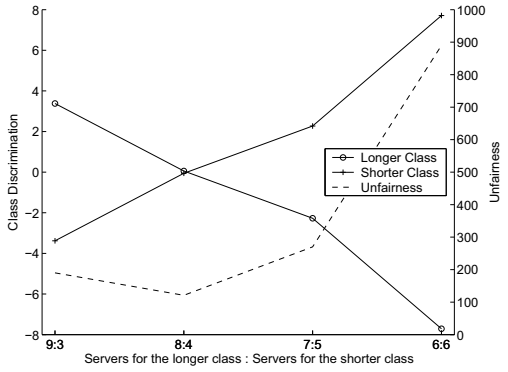


Figure 3: Unfairness and Class Discrimination in a 12 server Dedication System

The figure demonstrates that the discrimination of a class increases with the number of servers allocated to it. It also demonstrates that equal assignment of servers negatively (and drastically) discriminates class 1 and that the best operation point (smallest unfairness value and smallest absolute values of class discriminations) is to assign the servers proportionally to the mean service requirement of the class. Nonetheless, the class discrimination experienced at the optimal operation point (8:4) is *not exactly zero*, resulting from the slight differences in behavior between the 8 server system and the 4 server system.

For comparison we also evaluate a system where the two classes mix together and share the 12 servers (under a single FCFS queue). In this system the unfairness becomes very small (13.773 compared to 121 in the dedicated server system), suggesting that this is a more fair strategy. As for class discrimination, its absolute values are quite small (0.20995 and -0.20995). Note that the large jobs (class 1) now enjoy positive discrimination while the short jobs (class 2) suffer from negative discrimination, as expected from Section 3.1.

We repeated this examination for uniformly distributed service times and found similar results.

We now move on to another interesting scenario, the *rush hour scenario*. In some cases customers appear over a short period of time, the "rush hour", and are served from that moment until all customers leave the system. One example for such a system is a computerized call centers or an Internet Web server, which expects a large influx of customers arriving concurrently, due to a major TV advertisement. Such a system may be expecting two customer classes, previously registered members and new members, with different service times. A second example is the queue for the restrooms in theaters, at the beginning of the theater break. For simplic-

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

ity we assume that all customers arrive concurrently.

For the sake of presentation, and to obtain tractable results, we consider a simple deterministic example. We assume a total number of 12 servers and $12j$ customers in each class, j even. The service times of class 1 and class 2 are 1 and 2 time units respectively. We consider two intuitive and common server assignment policies: 1) Proportional to the service length, that is 4 and 8 servers to class 1 and 2, respectively, and 2) Equal, namely 6 servers to each class.

We track the system in the proportional case along time slots of one time unit. Since $12j$ jobs of each type are present at time zero, the number of slots is $3j$. 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 1 (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 1 is $12j$, and that there are $12j$ customers, the class discrimination of class 1 is given by

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

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

For the equal allocation case a similar analysis yields the overall warranted service of class 1 to be: $12\sum_{i=1}^{j} \frac{2i-1}{j+3i-1} + \frac{2i}{3i+j}$, which yields the class discrimination:

$$\mathbb{E}\{D_{(1)}\} \approx 1 - \frac{1}{9j}\Big(-12 + 12j - 2j\ln 4 + 2j\ln\frac{3+j}{j}$$
$$+ (1+2j)\big(\ln(2+j) - \ln(4j-1)\big)\Big)$$

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

The analysis reveals that under our fairness model, neither proportional assignment nor equal assignment is most 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 $\mathbb{E}\{D_{(1)}\} \approx 0.06$.

5. CONCLUDING REMARKS

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 metric called class-discrimination, which is a variant of RAQFM, 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 importance of these results is that it is perhaps the first attempt to evaluate the fairness aspects of these priority mechanisms and conduct it in quantitative manner. The results derived and the techniques developed can be used for conducting fairness evaluation of such practical systems in a variety of applications.

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. These call for future research. In particular, fairness aspects of prioritization and resource dedication via alternative fairness approaches, such the ones mentioned in Section 1.3, is a subject of current research.

Acknowledgements

This work was supported in part by the Israeli Ministry of Science and Technology, grant number 380-801, and by EURO-NGI network of excellence.

6. REFERENCES

[1] I. Adan and J. Resing. Queueing theory. Online book, 2001.

[2] B. Avi-Itzhak, E. Brosh, and H. Levy. SQF: A slowdown queueing fairness measure. *Perform. Eval.*, 64(9–12):1121–1136, 2007.

[3] B. Avi-Itzhak, H. Levy, and D. Raz. Quantifying fairness in queueing systems: Principles, approaches and applicability. *Probability in the Engineering and Informational Sciences (PEIS)*, 22(4), 2008.

[4] S. L. Brumelle. On the relation between customer and time averages in queues. *Journal of Applied Probability*, 8(3):508–520, September 1971.

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

[6] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queueing algorithm. *Internetworking Research and Experience*, 1:3–26, 1990.

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

[8] A. G. Greenberg and N. Madras. How fair is fair queueing? *Journal of the ACM*, 3(39):568–598, 1992.

[9] J. M. Jaffe. Bottleneck flow control. *IEEE Transactions on Communications*, 29(7):954–962, July 1981.

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

[11] F. P. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8:33–37, 1997.

[12] A. A. Kherani and A. Kumar. Stochastic models for throughput analysis of randomly arriving elastic flows in the internet. In *Proceedings of IEEE INFOCOM '02 21st Annual Joint Conference of the IEEE Computer*

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344

and Communications Societies, volume 2, pages 1014–1023, June 2002.

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

[14] R. A. Marie. Calculating equilibrium probabilities for $\lambda(n)/c_k/1/n$ queue. In *Proceedings of Performance '80*, pages 117–125, Toronto, May 1980.

[15] 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.

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

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

[18] A. Rafaeli, E. Kedmi, D. Vashdi, and G. Barron. Queues and fairness: A multiple study experimental investigation. Technical report, Faculty of Industrial Engineering and Management, Technion. Haifa, Israel. Under review, 2003.

[19] J. Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[20] D. Raz. *Quantifying Job Fairness in Queueing Systems*. PhD thesis, School of Computer Science, Tel-Aviv University, 2008.

[21] D. Raz, B. Avi-Itzhak, and H. Levy. Fairness considerations in multi-server and multi-queue systems. In *Proceedings of Valuetools First International Conference on Performance Evaluation Methodologies and Tools*, Pisa, Italy, October 2006. Article 39.

[22] 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).

[23] G. Sabin, G. Kochhar, and P. Sadayappan. Job fairness in non-preemptive job scheduling. In *Proceedings of the 2004 International Conference on Parallel Processing (ICPP-04)*, pages 186–194, Montreal, Quebec, Canada, August 2004.

[24] W. Sandmann. A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement. In R. Sabella, editor, *Proceedings of the 1st Euro NGI Conference on Next Generation Internet Networks-Traffic Engineering (NGI2005)*, pages 106–113, Rome, Italy, April 2005.

[25] L. E. Schrage and L. W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14:670–684, 1966.

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

[27] A. Wierman. Fairness and classifications. *Perform. Eval. Rev.*, 34(4):4–12, 2007. *Special Issue on New Perspectives in Scheduling*.

[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.

[29] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to higher moments of conditional response time. In *Proceedings of ACM Sigmetrics 2005 Conference on Measurement and Modeling of Computer Systems*, pages 229–239, Banff, Alberta, Canada, June 2005.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4344
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4344