# A Resource Allocation Queueing Fairness Measure: Properties and Bounds

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

**Hanoch Levy**
School of Computer Science
Tel-Aviv University, Tel-Aviv, Israel
hanoch@cs.tau.ac.il

**David Raz**
School of Computer Science
Tel-Aviv University, Tel-Aviv, Israel
davidraz@post.tau.ac.il

May 25, 2005

## Abstract

Fairness is an inherent and fundamental factor of queue service disciplines in a large variety of queueing applications, ranging from airport and supermarket waiting lines to computer and communication queueing systems. Recent empirical studies show that fairness is highly important to queueing customers in actual situations. Despite this importance, queueing theory has devoted very little effort to this subject and an agreed upon measure for evaluating the fairness of queueing systems does not exist. In this work we study a newly proposed Resource Allocation Queueing Fairness Measure (RAQFM). The measure, first introduced in Raz et al. (2004d), is built under the understanding that a widely accepted measure must adhere to the common sense intuition of researchers as well as practitioners and customers, and must also be based on widely accepted principles of social justice. We analyze the properties of RAQFM and provide bounds for its values. Both of these serve to intuitively understand the measure and provide confidence in it. The analysis shows that the measure properly reacts to both *customer seniority* and *customer service time*, and thus appeals to one's intuition. The bounds provide a scale of reference on the measure. An Additional property of the measure, namely "locality of reference", and how it yields to analysis, are discussed.

Subject classifications: Queues: quantification of job fairness in queueing systems
Area of review: Stochastic Models

# 1 Introduction

Queueing systems are encountered in a wide variety of applications such as supermarkets, airports, banks, public offices, computer systems, communication systems, web services, call centers, and many others. Queueing Theory has been used for nearly a century to study the performance of such systems and how to operate them efficiently.

Why are ordered queues used in all these real life situations? Perhaps the major reason for using an *ordered queue* at all is to provide fair service to the customers; in this sense one can view a queue as a "fairness management facility". Furthermore, empirical evidence to the importance of fairness in queues was provided recently in Rafaeli, Barron, and Haber (2002) and Rafaeli et al. (2003). Their work uses an experimental psychology approach to study the reaction of humans to waiting in queues and to various queueing and scheduling policies. These studies revealed that for humans waiting in queues, the issue of fairness is highly important, perhaps sometimes even more important than the duration of the wait.

The fairness factor associated with waiting in queues has been recognized in many works and applications; some of them are listed next. Larson (1987) in his discussion paper on the disutility of waiting, recognizes the central role played by 'Social Justice', (which is another name for fairness), and its perception by customers. This is also addressed by Rothkopf and Rech (1987) in their paper discussing perceptions in queues. Aspects of fairness in queues were discussed by quite a number of authors, including Palm (1953) that deals with judging the annoyance caused by congestion, Mann (1969) that discusses the queue as a social system and Whitt (1984) that addresses overtaking in queues.

Despite the importance of queue fairness, little has been published on how to quantify it. As a result, the issue of fairness is not generally understood, and widely agreed upon measures do not exist. Thus, the fairness of real applications cannot be evaluated and systems cannot be compared to each other. Some research exceptions are Gordon (1987), Avi-Itzhak and Levy (2004), Bender, Chakrabarti, and Muthukrishnan (1998), Bansal and Harchol-Balter (2001), and Wierman and Harchol-Balter (2003). In Gordon (1987) the number of "skips" and "slips" experienced in the queue by an arbitrary customer is proposed to reflect the queue injustice (and an analysis of this metrics was carried out for several systems). In Avi-Itzhak and Levy (2004) measures based on order of service have been devised. The slowdown (a.k.a. stretch, normalized response time) was proposed as a metric of unfairness in several works. In Bender, Chakrabarti, and Muthukrishnan (1998) the *max slowdown* is used as indication of unfairness. In Bansal and Harchol-Balter (2001) the *max mean slowdown* is used to evaluate the unfairness of the SRPT scheduling policy. In Wierman and Harchol-Balter (2003), the max mean slowdown is used as a criterion for evaluating whether a system is fair or unfair.

A large volume of literature exists on weighted fair queueing. However, that work is outside the scope of this paper. It deals with fairness to streams, fitting communications systems mainly, rather than with fairness to jobs, which is the subject of this paper.

In light of the importance of fairness to queue-based applications, the objective of this paper is to study a newly proposed methodology and a metric that can be applied to queueing systems and scheduling policies for evaluating their level of fairness. To properly devise such a method one should first ask what are the basic physical properties playing a role in queue fairness. To this end observe that the behavior of a queueing system is governed by two major physical factors, *job seniority* and *job service requirements* (the terms "job" and "customer" are used interchangeably throughout the paper). In every queueing analysis they serve, in the form of arrival times and service times, together with the server policy, to derive the system performance measures (e.g., expected delay). Thus, a complete fairness measure should account for both[1]. To demonstrate how seniority and service times affect fairness, consider the following daily-life scenario, taken from the supermarket queue setup: Mr. Short arrives to a supermarket queue with a couple of items and finds in front of him Mrs. Long with an overflowing cart. The question of whether it is fair to serve Short ahead of Long, and the dilemma associated with this question, is rooted in the contradicting physical factors of *seniority difference* (working to the benefit of Long) and *service requirement difference* (working to the benefit of Short). The prior recent work mentioned above focused on one of these physical factors. The "skip and slip" approach (Gordon (1987)) and order-of-service based measure (Avi-Itzhak and Levy (2004)) focus on the issue of seniority; the latter has a modification of the measure to account for service times. The slowdown approach (Wierman and Harchol-Balter (2003)) successfully captures the service time differences between jobs and provides interesting results regarding which policies are fair in this regard. However, that approach does not

[1] One may claim that in some applications, such as computer systems or call centers, the issue of seniority is not important and only size should matter, since the customers anyhow do not see each other. However, we believe that in most of these applications customers do care about seniority even if they do not see each other. For example, we believe that customers might be quite upset if they find out that their phone-accessed bank teller serves them in LCFS order.

account for seniority differences.

Very intriguing and drastically contradicting results may be obtained if only one of the factors is accounted for. Thus, a measure that accounts only for seniority differences, as shown in Avi-Itzhak and Levy (2004), will rank First-Come-First-Served (FCFS) as the most fair policy and Last-Come-First-Served (LCFS) as the most unfair policy. In contrast, a criterion that accounts only for service time differences, such as the criterion developed in Wierman and Harchol-Balter (2003), classifies Preemptive LCFS as always fair and FCFS as always unfair.

In this work we address a new measure that accounts both for *seniority differences* and for *service time differences*, and is convenient for analysts to work with. To achieve this, our approach focuses on the *server resources* and examines how fairly they are allocated to the jobs. The approach is called a Resource Allocation Queueing Fairness Measure (RAQFM) and was first introduced in Raz, Levy, and Avi-Itzhak (2004d). This measure is based on the basic ("axiomatic") belief, stemming from the widely accepted social justice principle of equally dividing the "pie", that at every epoch all jobs present in the system deserve an equal share of the server's attention ("pie"). This is the case with Processor Sharing (see analysis, as early as Kleinrock (1964, 1967), Coffman, Muntz, and Trotter (1970), followed by many others). Deviations from this principle are assumed to create customer discriminations (positive or negative). Accounting for these discriminations and summarizing them yields a measure of unfairness. Detailed description is given in Section 3 (after a short overview of some other work on fairness, given in Section 2).

The main objective of this work is to examine the basic properties of the RAQFM measure. We believe that three types of properties are desired for such a measure: 1) Agreeing with one's intuition in special cases, 2) Having computable bounds on the measure, and 3) Yielding to analysis. To address the first type of properties, note that a fairness measure is somewhat an "abstract" entity that is "hard to feel"; thus, the evaluation of the measure in simple and widely agreed upon cases, and the examination of how it fits one's intuition (queueing experts as well as "plain customers"), can assist in examining the credibility of the measure and building confidence in it. Such confidence is important for the measure to be used in evaluating complex and subtle cases. The first case for which we examine RAQFM is where preemption is not allowed and all service times are identical, either deterministically or stochastically, that is, only seniority matters. In this case we show that serving a senior ahead of a junior increases fairness, which fits with intuition. The second case is where all arrival times are identical (thus only service time matters). In this case we show that serving a short job ahead of a long job increases fairness, again fitting one's intuition. Third, we show that Processor Sharing (PS) is the most fair policy and that this optimality is unique to PS and its precise imitators. These properties are derived in Section 4.

The importance of the second type of properties, namely bounds, is to provide some scale of reference. Such scale of reference is useful when the measure is used to evaluate a system and some intuitive meaning of the fairness numbers is needed. Bounds on the discrimination values according to RAQFM are derived in Section 5; these bounds were stated in Raz et al. (2004d) and are first proved in this work.

The third desired property of a measure is to yield to analysis. To this end first note that an important property of RAQFM is that it is based on first accounting for the *individual discriminations* attributed to each job in the system and then summarizing them; this allows one to use RAQFM for several purposes: i) Measuring *individual* job discrimination in a *specific sample path*, ii) Evaluating the overall unfairness of *a scenario* (a sample path), and iii) Evaluating the unfairness of systems and service policies (by evaluating the unfairness in steady state). These allow practitioners and customers to get a feel of the fairness they encounter in the system. Second, in Section 6 we demonstrate a method by which RAQFM can be derived for Markovian systems in steady state. This is demonstrated for the FCFS service discipline.

An additional important property of RAQFM is that the discrimination function used by RAQFM possess a "locality of reference" property, namely, that its variance over all customers is identical to its variance over the customers of a busy period. This property is important for proper fairness evaluation. In Section 7 we briefly discuss this property, whose full analysis is postponed to a forthcoming paper, due to lack of space.

Lastly (Section 8) we provide numerical results that further demonstrate the sensitivity of RAQFM to service time and seniority. The results seem to fit intuition and thus provide additional confidence in the measure.

Concluding remarks are given in Section 9.

## 2 Short Overview of Some Other Work on Fairness

One area where several fairness measures were proposed is flow control. Two well known notions in this area are those of *Max-Min Fairness* (Starting with Jaffe (1981) and used by many afterwards) and Proportional Fairness (Kelly (1997)). These notions deal with fair allocation of *rates* (or *bandwidth*) to customers, and are not applicable to fair scheduling of customers in a queue.

Another related area where there has been research on the matter of fairness is fair queueing. The measures mostly used in this area are *Absolute Fairness Bound* (AFB) and *Relative Fairness Bound* (RFB). AFB (first used probably in Greenberg and Madras (1992)) is based on the maximum difference between the service received by a flow under the discipline being measured, and that it would have received under the ideal PS policy. As AFB is frequently hard to obtain (see Keshav (1997, ch. 9 pp. 209-261)) RFB was proposed (first used probably by Golestani (1994)), based on the maximum difference between the service received by any two flows under the policy being measured. See Zhou and Sethu (2002) for relations between AFB and RFB. Both of these measures were originally meant to be used in studying flows and not specific jobs, although they can be applied to jobs as well. For example, for AFB one can compute the maximum difference between the departure time of each job and the departure time it would have received under PS. However, when either of these measures is used for evaluating job fairness, the following emerges:

1. If job sizes are unbounded, these measures are unbounded (i.e. infinitely unfair) for all non-preemptive policies. In fact, the tightest bound possible for any non-preemptive policy is the size of the largest job (achieved by the Fair Queueing policy proposed by Demers, Keshav, and Shenker (1989, 1990)).

2. Even if job sizes are bounded, it is easy to see that these measures do not differentiate between many non-preemptive service policies. For example, both FCFS and LCFS are equally and infinitely unfair, and so are Shortest Job First (SJB), Longest Job First (LJB) and Random Order of Service (ROS).

Both cases above imply that these measures, based on a maximal-difference approach are not accurate enough to differentiate between many popular scheduling policies which drastically differ from each other.

A similar criterion, with similar properties, was suggested by Friedman and Henderson (2003). According to that criterion, a protocol $p$ is considered fair if it weakly dominates PS, namely no job completes later under $p$ than under PS, on any sample path. This criterion is similar to AFB in that it compares the protocol against PS, and it considers the worst case scenario, though it only classifies the protocol as fair or unfair. Again, all non-preemptive policies are unfair, as well as most preemptive policies.

Another work worth mentioning is Wang and Morris (1985), where the Q-factor is proposed for measuring the performance of load sharing algorithms. It measures the performance, relative to multi-server FCFS, as observed by the customer source treated worst, under the worst possible combination of loads. While the measure is mainly introduced to detect inefficiencies in the load sharing algorithm it also has some fairness aspects.

# 3 Introducing RAQFM in a Single Server System

## 3.1 Model and Notation

Consider a queueing system with one server. The system is subject to the arrival of a stream of customers, $C_1, C_2, \dots$, who arrive at the system at this order. Let $a_i$ and $d_i$ denote the arrival and departure epochs of $C_i$ respectively. Let $s_i$ denote the service requirement (measured in time units) of $C_i$. A specific series of values $\{a_i\}_{i=1,2,\dots,L}$ is called an *arrival pattern*. A specific series of values $\{a_i, s_i\}_{i=1,2,\dots,L}$ is called an *arrival and service pattern*. A specific series of values $\{a_i, s_i, d_i\}_{i=1,2,\dots,L}$ is called a *scenario*.

At each epoch $t$ the server grants service at rate $s_i(t) \geq 0$ to $C_i$. Let $N(t)$ denote the number of customers in the system at epoch $t$. The system is work-conserving, i.e. $\int_{a_i}^{d_i} s_i(t)dt = s_i$. The server has a service rate of one unit and is non-idling, i.e. $\forall t, N(t) > 0 \Rightarrow \sum_i s_i(t) = 1$.

All customers are "born equal", and thus no weights are assigned to them. In an ongoing research we deal with a weighted version of the measure, meant to be used in cases where customers are not equal.

## 3.2 Individual Customer Discrimination

The fundamental principle underlying RAQFM is the belief that at every epoch $t$, all customers present in the system deserve an equal share of the system resources. This principle implies that the share of the server resources a customer deserves at $t$ is simply given by $1/N(t)$. We call this quantity the *warranted service rate* of $C_i$ at epoch $t$, and denote it $R_i(t)$. Integrating this for $C_i$ yields $R_i \overset{def}{=} \int_{a_i}^{d_i} dt/N(t)$, the *warranted service* of $C_i$. The *(overall) discrimination* of $C_i$, denoted $D_i$, is the difference between the warranted service and the granted service, i.e.

$$D_i = s_i - R_i = s_i - \int_{a_i}^{d_i} dt/N(t). \tag{1}$$

A positive (negative) value of $D_i$ means that a customer receives better (worse) treatment than it fairly deserves, and therefore it is *positively (negatively) discriminated.*

An alternative way to define $D_i$ is to define the *discrimination rate* of $C_i$ at epoch $t$,

$$\delta_i(t) \overset{def}{=} s_i(t) - 1/N(t), \tag{2}$$

and then the overall discrimination of $C_i$ is:

$$D_i = \int_{a_i}^{d_i} \delta_i(t)dt. \tag{3}$$

An important property of this measure is that it obeys, for every non-idling work-conserving system, and for every $t$: $\sum_i \delta_i(t) = 0$, that is, every positive discrimination is balanced by negative discrimination. This results from the fact that when the system is non-empty $\sum_i s_i(t) = 1$ (due to non-idling) and $\sum_i R_i(t) = N(t)(1/N(t)) = 1$. An important outcome of this property is that if $D$ is a random variable denoting the discrimination of an arbitrary customer when the system is in steady state, then $E[D] = 0$, namely the expected discrimination is zero. A complete proof is given in Raz, Avi-Itzhak, and Levy (2004b).

### 3.3 Unfairness of a Scenario

To evaluate the unfairness of a scenario one can compute the set of individual discriminations $D_i, i = 1, \ldots, L$ using Eq.(1) or Eq.(3).

One would then choose some summary statistics measure over the values $D_i$. Since fairness inherently deals with differences in treatment of customers a natural choice is the statistical variance of customer discrimination. Since the average of $D_i$ is zero, this equals the statistical second moment, that is $\frac{1}{L}\sum_{i=1}^{L}(D_i)^2$. We denote this measure $F_{D^2}$. Another optional measure is the average distances $\frac{1}{L}\sum_{i=1}^{L}|D_i|$ (denoted $F_{|D|}$) or the average absolute value of negative discrimination (denoted $F_{D<0}$).

### 3.4 System Measure of Unfairness

To measure the unfairness of a system and of a service policy across all customers, that is, to measure the *system unfairness*, one would choose some summary statistics measure over $D$, where $D$ is a random variable denoting the discrimination of an arbitrary customer when the system is in steady state.

Again, one can naturally choose the variance of customer discrimination, and since $E[D] = 0$, this equals the second moment. Similarly, another option is the mean of distances $E[|D|]$. We use $F_{D^2}, F_{|D|}, F_{D<0}$ to denote these too, the meaning being obvious from the context. Throughout this paper, the term “unfairness” refers to $F_{D^2}$ since the paper focuses on this measure. In some instances we also mention $F_{|D|}$.

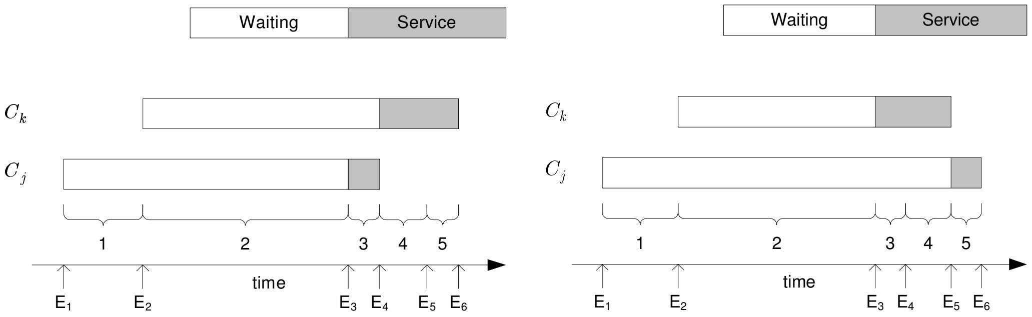
*Remark* 3.1 *(The connection between scenario fairness and system fairness).* Note that system fairness deals with the expectation over all scenarios, while scenario fairness deals with the realization of a specific scenario. Thus, if for all arrival and service patterns policy $\phi_1$ is more fair than policy $\phi_2$ then this property is true also for the system unfairness.

## 4 Properties of RAQFM

### 4.1 Reaction to Differences in Seniority

In this section we show that for a commonly encountered class of policies RAQFM reacts well to seniority differences. We do this by showing that in the special case where service times are identical, either deterministically or stochastically, RAQFM “prefers” serving in order of seniority. In the deterministic case we show that providing preferential service to senior customers yields lower unfairness values for every sample path. That is, RAQFM assigns lower unfairness values to schedules in which senior customers get preferential service over junior customers. In the stochastic case, we show that providing preferential service to senior customers yields lower value of expected unfairness. In both cases we deal with non-preemptive policies exclusively. More specifically, define $\Phi$ to be the class of non-preemptive, non-divisible service polices (i.e. service policies where once the server started serving a customer it will not stop doing so until the customer's service requirement is fulfilled, and at most one customer is served at any epoch), where the scheduler does not know the actual values of the service times, or does not account for them in the service decisions. We deal with policies in $\Phi$.

Consider two customers $C_j$ and $C_k$ that are adjacently served, with arrival times $a_j < a_k$ and service requirements $s_j, s_k$. Observe the two possible scenarios: In scenario (a), (Figure 1(a), seniority preserving schedule), the order of seniority is preserved, i.e. $C_j$ is served before $C_k$. In scenario (b), (Figure 1(b), seniority violating schedule), the order of service of $C_j$ and $C_k$ is interchanged and thus the order of seniority is violated. For every other customer, the arrival time, service requirement, and departure time are the same across the two scenarios. We assume that both of the schedules are *possible* , i.e. that $C_j$ and $C_k$ reside in the queue together for some time, and thus are interchangeable.



(a) Seniority Preserving Schedule

(b) Seniority Violating Schedule

Figure 1: Two Adjacently Served Customers

We now present some general notation that is used in this section.

Let $D_i^a$ denote the discrimination of $C_i$ under scenario (a) and $D_i^b$ the discrimination under scenario (b). Let $F_{D^2}^a$, $F_{D^2}^b$ denote the unfairness in the respective scenarios, and let $N^a(t)$, $N^b(t)$ denote the number of customers in the system at epoch $t$, respectively.

Observe that for every customer $C_i$, $D_i$ is determined by $a_i$, $s_i$, $d_i$, and $N(t)$ in the interval $(a_i, d_i)$, $i = 1, 2, \ldots, L$. The interchange of $C_j$ and $C_k$ affects only $d_j$, $d_k$, and $N(t)$.

Let $\tilde{F}$ denote the total unfairness to all customers other than $C_j$ and $C_k$, and let $\hat{F}$ denote the total unfairness to $C_j$ and $C_k$. Then $F_{D^2} = \hat{F} + \tilde{F}$.

Define $\Delta F_{D^2}$, $\Delta\hat{F}$, and $\Delta\tilde{F}$ to be the change due to the interchange in the values of

$F_{D^2}$, $\hat{F}$, and $\tilde{F}$ respectively. Then

$$\Delta F_{D^2} = \Delta\hat{F} + \Delta\tilde{F} = \frac{1}{L}\left[(D_j^b)^2 + (D_k^b)^2 - (D_j^a)^2 - (D_k^a)^2\right] + \frac{1}{L}\left[\sum_{i \neq j,k}^{L}(D_i^b)^2 - \sum_{i \neq j,k}^{L}(D_i^a)^2\right]. \quad (4)$$

We assume now that $s_j = \tau_1$ and $s_k = \tau_2$, and denote $\tau_{min} = \min(\tau_1, \tau_2)$ and $\tau_{max} = \max(\tau_1, \tau_2)$. We note that the time interval from the arrival of $C_j$ ($a_j$ in Figure 1) and $\max(d_j, d_k)$ (namely until both $C_j$ and $C_k$ depart) is of equal length in the two scenarios. We divide this interval into five sub-intervals $(E_i, E_{i+1})$, $i = 1, 2, 3, 4, 5$, where

1. $E_1 = a_j$
2. $E_2 = a_k$
3. $E_3$ is the first point in time where service, to either $C_j$ or $C_k$, starts.
4. $E_4 = E_3 + \tau_{min}$
5. $E_5 = E_3 + \tau_{max}$
6. $E_6 = \max(d_j, d_k)$

We note that $N^a(t) = N^b(t)$ for every $t$ except in the fourth interval $(E_4, E_5)$ (specifically, in the fourth interval $N^b(t) = N^a(t) + 1$ if $\tau_1 < \tau_2$ and $N^b(t) = N^a(t) - 1$ if $\tau_1 > \tau_2$). Therefore, the warranted service rate over interval $i$ (of any customer), $R^a_{(i)}$ and $R^b_{(i)}$ in scenario (a) and (b) respectively, is the same except for $i = 4$. Namely $R^a_{(i)} = R^b_{(i)}$, $i = 1, 2, 3, 5$ always and $R^a_{(4)} \neq R^b_{(4)}$ for $\tau_1 \neq \tau_2$.

### 4.1.1 Deterministically Equal Service Requirement

We start our discussion with the case where the two adjacently served customers have equal service requirement. We do not impose any restriction on the service requirements of other customers.

**Theorem 4.1 (Preference of Seniority Between Adjacently Served Customers with Equal Service Requirement).** *Let $C_j$ and $C_k$ be adjacently served customers, where $a_j < a_k$ and $s_j = s_k$. Then the scenario unfairness (measured either as $F_{D^2}$ or as $F_{|D|}$) of the seniority preserving schedule (a) is smaller than that of the seniority violating schedule (b), for every arrival and service pattern.*

*Proof.* We prove the theorem for $F_{D^2}$. The proof for $F_{|D|}$ is similar, and is not shown for conciseness.

Using Eq.(4), we need to prove that if $s_j = s_k$ then $\Delta F_{D^2} \geq 0$. Suppose $\tau_1 = \tau_2 = \tau$. In this case $E_4 = E_5$ (the fourth interval is of length zero) and $D_i^b = D_i^a, \forall i \neq j, k$. Therefore $\Delta \tilde{F} = 0$ and, from Eq.(4), we have

$$\Delta F_{D^2} = \Delta \hat{F} = \frac{1}{L}\left[(D_j^b)^2 + (D_k^b)^2 - (D_j^a)^2 - (D_k^a)^2\right]$$
$$= \frac{1}{L}\Big[\big(\tau - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a + R_{(5)}^a)\big)^2 + \big(\tau - (R_{(2)}^a + R_{(3)}^a)\big)^2$$
$$- \big(\tau - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a)\big)^2 - \big(\tau - (R_{(2)}^a + R_{(3)}^a + R_{(5)}^a)\big)^2\Big] = \frac{2}{L} R_{(1)} R_{(5)} > 0. \quad (5)$$

Note that we use the fact that for $\tau_1 = \tau_2$ we have $R_{(i)}^a = R_{(i)}^b, i = 1, 2, 3, 4, 5$. □

We also note that if $a_j = a_k$ (the two customers arrive concurrently), $\Delta F_{D^2} = 0$, as expected.

**Theorem 4.2 (Preference of Seniority Between Any Two Customers with Equal Service Requirement).** *Let $C_j$ and $C_k$ be any two customers, where $a_j < a_k$ and $s_j = s_k$. Then the scenario unfairness (measured either as $F_{D^2}$ or as $F_{|D|}$) of the seniority preserving schedule (a) is smaller than that of the seniority violating schedule (b), for every arrival and service pattern.*

*Proof.* The same proof holds, where the fifth interval now includes the interval between the epoch where the first customer (either $C_j$ or $C_k$) departs, and the other customer begins service. □

**Theorem 4.3 (Fairness of FCFS and LCFS for G/D/1).** *If the service requirements of all customers are identical (e.g. in the $G/D/1$ model), then for every arrival pattern, FCFS is the least unfair service policy in $\Phi$ and LCFS is the most unfair one. In other words, FCFS is the policy with the lowest system unfairness in $\Phi$, and LCFS is the one with the highest one.*

*Proof.* Assume for the contradiction that there exist an arrival pattern and a service policy $\phi \in \Phi$, $\phi \neq FCFS$, which is the least unfair policy in $\Phi$ for this arrival pattern. Then the order of service created by $\phi$ for this arrival pattern is different than the order of service created by FCFS, otherwise $\phi$ is indistinguishable from FCFS.

Given this arrival pattern and the order of service created by $\phi$, identify the first pair of customers which are adjacently served and are not served according to their order of arrival. Interchange the order of service between these two customers (which is certainly possible since the service policy is non-preemptive). According to Theorem 4.1, the result of this interchange is a decrease in the overall unfairness. Thus the resulting order of service is more fair than $\phi$, in contradiction to $\phi$ being the least unfair service policy for this arrival pattern.

A similar argument proves that LCFS is the most unfair policy in $\Phi$. □

*Remark* 4.1. From Remark 3.1 it follows that the properties stated in Theorem 4.3 for every arrival pattern, hold for system unfairness as well.

### 4.1.2 Stochastically Identical Service Requirement

**Theorem 4.4 (Preference of Seniority Between Adjacently Served Customers with Stochastically Equal Service Requirement).** *In a single server queueing system using a given discipline $\phi \in \Phi$ where the arrival process is independent of the service times assume that the service times of $C_j$ and $C_k$, denoted $S_j$ and $S_k$ respectively, are i.i.d and independent of the service times of all other customers. Let $\{a_i, s_i\}, i = 1, ..., L$, be an arrival and service pattern where $a_j < a_k$, $C_j$ and $C_k$ are served adjacently with $C_j$ being first and their order of service is interchangeable; denote this as possible scenario. Then, interchanging the order of service of $C_j$ and $C_k$ in all such possible scenarios will result in an increase in the expected unfairness to the $L$ customers.*

*Proof.* As in Theorem 4.1, the proof is given for $F_{D^2}$ and a similar proof can be given for $F_{|D|}$.

For $\tau_1 < \tau_2$, $\tau_1, \tau_2 > 0$, consider a possible scenario (a) from Figure 1 for the case $S_j = \tau_1 < S_k = \tau_2$. For this case we rewrite Eq.(4) as follows

$$\Delta(F_{D^2}|S_j = \tau_1, S_k = \tau_2) = \Delta(\hat{F}|S_j = \tau_1, S_k = \tau_2) + \Delta(\tilde{F}|S_j = \tau_1, S_k = \tau_2). \qquad (6)$$

For the case $S_j = \tau_2 > S_k = \tau_1$ (arrival times and service requirements of all customers except $C_j$ and $C_k$ being the same as in the previous case) we have

$$\Delta(F_{D^2}|S_j = \tau_2, S_k = \tau_1) = \Delta(\hat{F}|S_j = \tau_2, S_k = \tau_1) + \Delta(\tilde{F}|S_j = \tau_2, S_k = \tau_1). \qquad (7)$$

Since the service times of $C_j$ and $C_k$ are i.i.d., scenario (a) with $S_j = \tau_1, S_k = \tau_2$ is equally likely to possible scenario (b) with $S_j = \tau_2, S_k = \tau_1$. Further, the former is possible if and only if the latter is possible, since the scheduling decisions prior to the service of $C_j$ and $C_k$ are independent of their values $\forall \phi \in \Phi$. Therefore it is sufficient to show that $\Delta(F_{D^2}|S_j = \tau_1, S_k = \tau_2) + \Delta(F_{D^2}|S_j = \tau_2, S_k = \tau_1) > 0$ for the theorem to be true (from Theorem 4.1 we already have $\Delta(F_{D^2}|S_j = S_k) > 0$).

We have

$$\begin{aligned}
&\Delta(\hat{F}|S_j = \tau_1, S_k = \tau_2) + \Delta(\hat{F}|S_j = \tau_2, S_k = \tau_1) \\
&\quad = \frac{1}{L}\Big[\big(\tau_1 - (R^a_{(1)} + R^a_{(2)} + R^a_{(3)} + R^b_{(4)} + R^a_{(5)})\big)^2 + \big(\tau_2 - (R^a_{(2)} + R^a_{(3)} + R^b_{(4)})\big)^2 \\
&\qquad - \big(\tau_1 - (R^a_{(1)} + R^a_{(2)} + R^a_{(3)})\big)^2 - \big(\tau_2 - (R^a_{(2)} + R^a_{(3)} + R^a_{(4)} + R^a_{(5)})\big)^2 \\
&\qquad + \big(\tau_2 - (R^a_{(1)} + R^a_{(2)} + R^a_{(3)} + R^a_{(4)} + R^a_{(5)})\big)^2 + \big(\tau_1 - (R^a_{(2)} + R^a_{(3)})\big)^2 \\
&\qquad - \big(\tau_2 - (R^a_{(1)} + R^a_{(2)} + R^a_{(3)} + R^b_{(4)})\big)^2 - \big(\tau_1 - (R^a_{(2)} + R^a_{(3)} + R^b_{(4)} + R^a_{(5)})\big)^2\Big] \\
&\quad = \frac{2}{L} R^a_{(1)}(R^a_{(4)} + 2R^a_{(5)}) > 0, \qquad (8)
\end{aligned}$$

where $R^a_{(4)}$ and $R^b_{(4)}$, each appearing several time in the above relation, are $(R^a_{(4)}|S_j = \tau_1, S_k = \tau_2)$ and $(R^b_{(4)}|S_j = \tau_1, S_k = \tau_2)$, respectively, and $\tau_1 < \tau_2$. We also use the fact that $R^a_{(i)} = R^b_{(i)}, i = 1, 2, 3, 5$. To clarify the relation between the four lines of expressions appearing in Eq.(8) note that the first and second lines correspond to Figure 1(b) and Figure 1(a) respectively, and the third and fourth lines correspond to a symmetric scenario with $S_j = \tau_2, S_k = \tau_1$ $(\tau_1 < \tau_2)$. Also, note that the right hand side of the equation does depend on $\tau_1$ and $\tau_2$ via the $R$ values.

For $C_i$, $i \neq j, k$, scenario (a) with $S_j = \tau_1$ and $S_k = \tau_2$ is identical to scenario (b) with $S_j = \tau_2$ and $S_k = \tau_1$ and vice versa. Therefore $\Delta(\tilde{F}|S_j = \tau_1, S_k = \tau_2) + \Delta(\tilde{F}|S_j = \tau_2, S_k = \tau_1) = 0$ and thus $\Delta(F_{D^2}|S_j = \tau_1, S_k = \tau_2) + \Delta(F_{D^2}|S_j = \tau_2, S_k = \tau_1) > 0$.

From Remark 3.1 it follows that if the set of all possible such scenarios is not of measure zero, the expected unfairness will increase (and otherwise it will not change). □

**Theorem 4.5 (System Unfairness of FCFS and LCFS for G/G/1).** *Consider a single server system with arbitrary arrivals and where the service times are i.i.d random* *variables with arbitrary known distribution (e.g. G/G/1). Then FCFS is the service policy with the lowest systen unfairness in $\Phi$ and LCFS is the one with the highest system unfairness.*

*Proof.* The proof follows immediately from Corollary 4.1.2 and using an argument similar to the one used in Theorem 4.3.

For a given arbitrary service requirement distribution, assume, for the contradiction, that there exists an arrival pattern, and a service policy $\phi \in \Phi$, $\phi \neq FCFS$, which is the policy with the lowest expected unfairness in $\Phi$ for this arrival pattern. Then the order of service created by $\phi$ for this arrival pattern is different from the order of service created by FCFS, otherwise $\phi$ is indistinguishable from FCFS.

Given this arrival pattern and the order of service created by $\phi$, observe the first pair of adjacently served customers which are not served according to their order of arrival. Since the more senior of these customers is served earlier by $\phi$, one can interchange their service order. According to Corollary 4.1.2, the result of this interchange is a decrease in the expected unfairness for this arrival pattern. Thus the resulting order of service is more fair than $\phi$, in contradiction to $\phi$ having the lowest expected unfairness for this arrival pattern.

A similar argument proves that LCFS has the highest system unfairness in $\Phi$. □

*Remark* 4.2. Again, from Remark 3.1 it implies directly that the properties stated in Theorem 4.5 for every arrival pattern, hold for system unfairness as well.

## 4.2 Reaction to Differences in Service Requirement

In this section we show that RAQFM reacts well to service requirement differences. We demonstrate this in the case where arrival times of all customers are identical.

**Theorem 4.6 (Preference of Shorter Service Time, For Simultaneously Arriving Customers.).** *Let $C_i, \quad i = 1, \ldots, N$ be $N$ customers arriving simultaneously (i.e. $\forall i, a_i = a$) at an empty system. Assume that no arrivals occur between $a$ and the departure epoch of the last customer $a + \sum_1^N s_i$. Then, for any two customers $C_i, C_j$ such that $s_i < s_j$, it is more fair to serve $C_i$ before $C_j$.*

*Proof.* For simplicity of presentation and without loss of generality, assume that customer index follows the customer's service order, namely $C_i$ is served before $C_{i+1}$, $i =$

$1, 2, \ldots, N-1$. The discrimination experienced by the $n$-th customer served is

$$D_n = s_n - \sum_{i=1}^{n} \frac{s_i}{N-i+1} = s_n \frac{N-n}{N-n+1} - \sum_{i=1}^{n-1} \frac{s_i}{N-i+1} \tag{9}$$

The unfairness of the scenario is

$$\frac{1}{N} \sum_{n=1}^{N} \left( s_n \frac{N-n}{N-n+1} - \sum_{i=1}^{n-1} \frac{s_i}{N-i+1} \right)^2 \tag{10}$$

To evaluate Eq.(10) we first evaluate the terms involving $s_n^2$. These yield

$$s_n^2 \left( \frac{N-n}{N-n+1} \right)^2 + \sum_{i=n+1}^{N} \left( \frac{s_n}{N-n+1} \right)^2 = \frac{N-n}{N-n+1} s_n^2. \tag{11}$$

Next consider the terms in the sum involving $s_n s_k$, $n > k$. These yield

$$-2 \frac{s_n(N-n)}{N-n+1} \frac{s_k}{N-k+1} + \sum_{i=n+1}^{N} 2 \frac{s_n}{N-n+1} \frac{s_k}{N-k+1} = 0. \tag{12}$$

To summarize, the unfairness of the scenario, namely Eq.(10), reduces to

$$\frac{1}{N} \sum_{n=1}^{N} \frac{N-n}{N-n+1} s_n^2. \tag{13}$$

Note that $\frac{N-n}{N-n+1}$ is monotone decreasing in $n$. Thus, the unfairness increases if a customer with larger service requirement is served ahead of a customer with smaller service requirement. In other words, the service order with the lowest unfairness is the one where $\forall i < j, s_i \leq s_j$, and every deviation from this order yields a higher unfairness order. □

**Corollary 4.1.** *It follows immediately from Theorem 4.6 that for a scenario consisting of $N$ simultaneously arriving customers (and no other customers), the most fair service order is Shortest Job First (SJF) and the least fair service order is Longest Job First (LJF).*

*Remark* 4.3. The advantage of serving a shorter service time customer $C_i$ ahead of a a longer service time customer $C_j$, as in Theorem 4.6, holds when arrival times of *all* customers are identical, and does not necessarily hold when only two customers arrive simultaneously, say $a_i = a_j$. For example, consider the following arrival and service pattern

$$\{(a_i, s_i)\}_{i=1\ldots5} = \{(0,3),(1,1),(1,2),(3,1),(6,1000)\} \tag{14}$$

and compare the following service orders: (i) $1,2,3,5,4$ (ii) $1,3,2,5,4$. Note that $a_2 = a_3$ and $s_2 < s_3$. Nonetheless, the unfairness of the first order of service is roughly $\approx 83556$ while that of the second order is roughly $\approx 83528$, namely the second order is more fair.

### 4.3 Absolute Fairness of PS

**Theorem 4.7 (Zero Unfairness of PS).** *For any arrival and service pattern, a scheduling policy has zero unfairness if and only if the departure epochs of all customers are identical to those in PS.*

*Remark* 4.4 *(PS Imitators).* A policy can schedule its processing in a way that the departure epochs of all customers are identical to those in PS, even if the scheduling is not identical to PS at every epoch. We call such a policy a "PS Imitator". We conjecture that in order to execute PS imitation a scheduler must know all the exact service times and arrival epochs of the customers ahead of time.

*Proof of Theorem 4.7.* First, PS has zero unfairness from the simple fact that for PS $s_i(t) = 1/N(t)$ for every epoch $t$ and for every customer in the system at that epoch. Thus,

$$\delta_i(t) = s_i(t) - 1/N(t) = 0 \Rightarrow D_i = \int_{a_i}^{d_i} \delta_i(t)dt = 0 \Rightarrow F_{D^2} = E[D^2] = 0, \tag{15}$$

where the first equality is from Eq.(2), and the second is from Eq.(3). Second, to consider PS imitators, observe that given the arrival epochs $a_i$, each discrimination value, and therefore the unfairness, are functions only of the departure epochs $d_i$ and of $N(t)$. Thus, a scheduling policy that has departure epochs equal to that of PS for the same arrival and service pattern has the same discrimination values, and therefore the same unfairness of PS.

Third, we prove the other direction of the theorem by way of contradiction. Assume for the contradiction that there exist an arrival and service pattern and scheduling policy $\phi$, with departure epochs that are not equal to those of PS, and that the resulting scenario has zero unfairness. Observe the first departure that is different from a departure according to

PS, say the departure of $C_k$. Denote the departure epoch according to PS and according to $\phi$ by $d_k$ and $d'_k$ respectively, where $d_k \neq d'_k$. Denote the discrimination of $C_k$ according to PS and according to $\phi$ by $D_k$ and $D'_k$ respectively. From the assumption $E[(D')^2] = 0$ and thus we must have $D'_k = 0$. Denote the number of customers in the system at epoch $t$ according to PS and according to $\phi$ by $N(t)$ and $N'(t)$ respectively. We have from Eq.(1) and Eq.(15)

$$D_k = s_k - \int_{a_k}^{d_k} dt/N(t) = 0. \tag{16}$$

Suppose $d_k > d'_k$, then all departures up to $d'_k$ are the same for PS and for $\phi$, and therefore $\forall t < d'_k, N'(t) = N(t)$. Thus,

$$\begin{aligned} D'_k = s_k - \int_{a_k}^{d'_k} dt/N'(t) &= s_k - \int_{a_k}^{d'_k} dt/N(t) \\ &= s_k - \left( \int_{a_k}^{d_k} dt/N(t) - \int_{d'_k}^{d_k} dt/N(t) \right) = \int_{d'_k}^{d_k} dt/N(t) > 0, \end{aligned} \tag{17}$$

where the inequality results from the fact that $N(t) \geq 1$ in $(d'_k, d_k)$ since $C_k$ is in the system. Thus, the assumption is contradicted.

Now suppose $d_k < d'_k$, then all departures up to $d_k$ are the same for PS and for $\phi$, and therefore $\forall t < d_k, N'(t) = N(t)$. Thus,

$$\begin{aligned} D'_k = s_k - \int_{a_k}^{d'_k} dt/N'(t) &= s_k - \left( \int_{a_k}^{d_k} dt/N'(t) + \int_{d_k}^{d'_k} dt/N'(t) \right) \\ &= s_k - \left( \int_{a_k}^{d_k} dt/N(t) + \int_{d_k}^{d'_k} dt/N'(t) \right) = - \int_{d_k}^{d'_k} dt/N'(t) < 0, \end{aligned} \tag{18}$$

again contradicting the assumption. □

**Corollary 4.2 (Absolute Fairness of PS).** *PS (and PS imitators) are the unique most fair scheduling policies.*

## 5 Bounds of RAQFM

In this section we derive bounds on the discrimination and unfairness measured by RAQFM. Note that several of the theorems brought here were previously mentioned briefly and without proof in Raz et al. (2004d). We use this opportunity to bring the theorems in their entirety and with a full proof.

## 5.1 Bounds On Individual Discrimination

**Theorem 5.1 (Bounds on Individual Discrimination).** *For every scenario and every customer $C_i$, $-W_i/2 \leq D_i < s_i$, where $W_i$ is the waiting time of $C_i$. Both bounds are tight.*

*Proof.* For the upper bound we have from Eq.(1)

$$D_i = s_i - \int_{a_i}^{d_i} dt/N(t)dt < s_i, \quad 1 \leq N(t) < \infty, s_i > 0. \tag{19}$$

To calculate the lower bound we divide the interval $(a_i, d_i)$ into two sets of intervals:

- $T_S^i = \{t|N(t) = 1\}$
- $T_W^i = \{t|N(t) > 1\}$.

We denote the length of a set of intervals $X$ by $\|X\|$.

From Eq.(1)

$$D_i = s_i - \int_{T_S^i} dt/N(t) - \int_{T_W^i} dt/N(t) \geq s_i - \int_{T_S^i} dt - \int_{T_W^i} dt/2 = s_i - \|T_S^i\| - \|T_W^i\|/2, \tag{20}$$

where the inequality is due to $-1/N(t) \geq -1/2, \forall t \in T_W^i$.

Note that $\|T_S^i\| + \|T_W^i\| = d_i - a_i$, thus the minimum is achieved when $\|T_S^i\|$ is the largest.

To bound $\|T_S^i\|$ observe that when $N(t) = 1$, $C_i$ must be served. As the system is work conserving, a customer cannot be served more than his requested service time, and therefore $\|T_S^i\| \leq s_i$. Thus, the minimum of Eq.(20) is achieved when $\|T_S^i\| = s_i \Rightarrow \|T_W^i\| = d_i - a_i - s_i = W_i$. Therefore,

$$D_i \geq s_i - s_i - W_i/2 = -W_i/2. \tag{21}$$

To show tightness of the upper bound we let $N(t) \to \infty, a_i \leq t \leq d_i$ in Eq.(19), where $d_i - a_i$ is finite. To show tightness of the lower bound consider the last customer in a FCFS busy period, who encounters exactly one customer in the system upon arrival.

Note that a customer may encounter a negative discrimination of $-W_i/2$ whose value is unbounded, even if service times are all bounded. This occurs to a customer who arrives to a LCFS served system with a single customer (in service) and who encounters a (possibly unbounded) sequence of arrivals occurring exactly at service completion epochs. □

## 5.2 Bounds on System Fairness

**Theorem 5.2 (Bounds on Scenario Unfairness).** *For every scenario,*

$$0 \le F_{|D|} \le 2s_{max} \tag{22}$$

$$0 \le F_{D^2} < \frac{N}{2}(s_{max})^2, \tag{23}$$

*where $s_{max}$ is the maximal service requirement and $N$ is the number of customers in the scenario. The lower bounds are both tight. The upper bound for $F_{|D|}$ is tight.*

*Proof.* For ease of reading we choose the unit of time to be $s_{max}$ , assuming $s_{max} < \infty$.

The lower bounds, including their tightness, were shown in Corollary 4.3.

The upper bounds can be derived by maximizing $1/N \sum_{i=1}^{N} |X_i|$ and $1/N \sum_{i=1}^{N} (X_i)^2$ under the constraints:

$$-\frac{N-1}{2} \le X_i \le 1 \tag{24}$$

$$\sum_{i=1}^{N} X_i = 0. \tag{25}$$

The first constraint arises from the bounds on the individual discrimination. The second constraint expresses $E[D] = 0$ (see Section 3.2).

For both $F_{|D|}$ and $F_{D^2}$, when $N \ge 3$ one of the global maxima is achieved at $X_1 = -(N-1)/2, X_2 = -(N-3)/2, X_i = 1$ for $i > 2$ (other global maxima exist, for example, due to symmetry between the variables, or for $F_{|D|}$ every $0 \le X_1, X_2 \le N-1$ where $X_1 + X_2 = -(N-2)$). The maximum value for $F_{|D|}$ is

$$\max F_{|D|} = \frac{1}{N}\big((N-1)/2 + (N-3)/2 + (N-2)\big) = 2 - 4/N < 2, \tag{26}$$

and the maximum value for $F_{D^2}$ is

$$\max F_{D^2} = \frac{1}{N}\left(\left((N-1)/2\right)^2 + \left((N-3)/2\right)^2 + (N-2)1^2\right) = N/2 - 1 + 1/2N < N/2. \tag{27}$$

To prove the tightness of the upper bound of $F_{|D|}$, consider a scenario as follows. All customers in this busy period have a service requirement of 1 unit of time. The scenario starts with the simultaneous arrival of $N$ customers (say $C_1, C_2, \ldots, C_N$) at the empty system. The first customer to be served is $C_N$. As soon as $C_N$ finishes service, a new customer (say $C_{N+1}$) joins the system and gets served ahead of $C_1, \ldots, C_{N-1}$. Just prior to the service completion of $C_{N+1}$, $C_{N+2}$ arrives and gets served ahead of $C_1, \ldots, C_{N-1}$, and so on, until $C_{N+M-1}$ is served. At the service completion of $C_{N+M-1}$, the first $N-1$ customers are served together using a processor sharing policy, and all leave the system $N+M-1$ units of time after the beginning of the scenario. Analyzing the above scenario we find $M$ customers with a positive discrimination of $1-1/N$ and $N-1$ customers with negative discrimination of $1 - M/N - (N-1)/(N-1) = -M/N$. The total unfairness is therefore $1/(M+N-1)(M(N-1)/N + (N-1)M/N) = 2M(N-1)/(N(M+N-1))$. Taking the limiting case $M \gg N \to \infty$, we get $F_{|D|} \to 2$. □

# 6 Computing RAQFM for Markovian Models

In this section we demonstrate how the system measure of RAQFM can be computed for Markovian models.

## 6.1 Analysis of the Single-Server FCFS System

To demonstrate one way of analyzing the system fairness we provide the analysis of the FCFS service M/M/1 system, with arrival rate $\lambda$ and mean service length $1/\mu$. This analysis was done in a slightly different approach in Raz, Levy, and Avi-Itzhak (2004d). It is provided here in order to demonstrate and clarify the methodology.

Let us consider a tagged customer $C$. For an arbitrary epoch let $a$ and $b$ denote the number of customers in the queue which are ahead of $C$ and behind $C$, respectively. Due to the memoryless properties of the system, the state $(a, b)$ (which we call the customer state) captures all that is needed for predicting the future discrimination of $C$.

The number of customers in the system at an epoch where $C$ observes the state $(a, b)$

is $a+b+1$. The discrimination rate at that state, denoted $\delta(a,b)$, is given by:

$$\delta(a,b) = \begin{cases} -\frac{1}{a+b+1} & a > 0, \\ 1 - \frac{1}{b+1} & a = 0. \end{cases} \tag{28}$$

The customer state will remain unchanged until the next customer arrival or departure occur. The duration until either of those events is exponentially distributed and its moments are

$$t^{(1)} = \frac{1}{\lambda+\mu}, \qquad t^{(2)} = \frac{2}{(\lambda+\mu)^2} = 2(t^{(1)})^2. \tag{29}$$

Since these moments are independent of the state we can denote them $t^{(1)}$ and $t^{(2)}$.

Let $D(a,b)$ be a random variable denoting the discrimination experience by $C$ through a walk starting at state $(a,b)$ and ending when $C$ leaves the system. Let $d(a,b)$ and $d^{(2)}(a,b)$ denote the first and second moment of $D(a,b)$, respectively.

Assume $C$ is in state $(a,b)$ at some epoch. The state of $C$ can change by one of the following events:

1. A customer arrives into the system. The probability of this event is $\tilde{\lambda} = \lambda/(\lambda+\mu)$. Afterwards $C$'s state will change to $(a, b+1)$.

2. A customer leaves the system. The probability of this event is $\tilde{\mu} = \mu/(\lambda+\mu)$. If $C$ is not being served $(a \neq 0)$ $C$'s state will change to $(a-1,b)$; else $C$ will leave the system.

This leads to the following recursive expression:

$$d(a,b) = t^{(1)}\delta(a,b) + \tilde{\lambda} d(a,b+1) + \begin{cases} \tilde{\mu} d(a-1,b) & a > 0, \\ 0 & a = 0. \end{cases} \tag{30}$$

Similarly, the equations for $d^{(2)}(a,b)$ are

$$d^{(2)}(a,b) = t^{(2)}(\delta(a,b))^2 + \tilde{\lambda} d^{(2)}(a,b+1) + 2t^{(1)}\delta(a,b)\tilde{\lambda} d(a,b+1) +$$
$$\begin{cases} \tilde{\mu} d^{(2)}(a-1,b) + 2t^{(1)}\delta(a,b)\tilde{\mu} d(a-1,b) & a > 0, \\ 0 & a = 0. \end{cases} \tag{31}$$

These expressions can be used, via numerical recursion, to compute the values of $d^{(2)}(a,b)$ to any desired accuracy.

We can now compute the system unfairness. Let $k$ be the number of customers seen by $C$ upon arrival into the system. Then

$$F_{D^2} = \sum_{k=0}^{\infty} p_k d^{(2)}(k, 0), \tag{32}$$

where $p_k = (1-\rho)\rho^k$ is the steady state probability of encountering $k$ customers in the system.

### 6.2 Analysis of Other Markovian Models

A methodology for analyzing a general Markovian system is provided in (Raz, , Levy, and Avi-Itzhak (2004a)).

This type of methodology, with various adaptations, was used to analyze the M/M/1 system with other service disciplines (Raz, Levy, and Avi-Itzhak (2004d)), M/M/1 queues with prioritization (Raz, Avi-Itzhak, and Levy (2004b)), M/M/r ($r > 1$) systems with a common queue and multiple servers (Raz, Avi-Itzhak, and Levy (2004c)) and queueing systems with Poisson arrivals and Coxian service time distributions (Brosh, Levy, and Avi-Itzhak (2004)).

## 7 Locality of Reference

One issue related to definition of fairness is the *locality of reference.* It may be argued that a customer's perception of the level of fairness is determined by how he is treated in comparison to other customers who, by getting higher or lower service preference, may impact his preference. These are customers who compete with him, locally in time, for the resources of the server. In a non-idling steady-state single-server system the largest possible group of reference is determined by the busy period in which a customer is being served. As shown in Raz, Levy, and Avi-Itzhak (2005) service preference of all customers in a busy period can be impacted by all other customers of the busy period. Customers served in two different busy periods cannot impact each other's service preference. A variance that is computed according to this principle is called *local-reference variance.*

In contrast, it may also be argued that the customer's perception of the level of fairness is determined by comparison to all customers ever served in the system (*globality of reference*). A variance that is computed according to this principle is called *global-reference variance.*

The RAQFM measure addresses both the locality of reference and the globality of reference approaches, since the values it computes under both approaches are always identical to each other. This is demonstrated in the following example. Consider a system where all customers have service time of one unit, and arrivals occur in bulks of either 2 or 4 customers in a bulk. Assume also that arrivals occur at constant distance of 6 time units, and thus each busy period consists of one bulk exactly. Suppose that the system is served in the Random Order of Service (ROS) discipline. The RAQFM value of customers served in the short busy period is 0.25 and in the long one it is 0.479. The probabilities of being served in a short and in a long busy periods are 1/3 and 2/3 respectively. Using the local-reference variance approach we get for RAQFM $F_{D^2} = (1/3)0.25 + (2/3)0.479 = 0.403$. Nontheless, since $E(D) = 0$ (zero-sum property) this is also the value of the global-reference variance.

In contrast, suppose one uses the waiting time variance as the unfairness measure (as opposed to the discrimination variance). In such a case the values of this measure for the short and long busy periods are 0.25 and 1.25 respectively. The local-reference variance unfairness measure value is then 0.917 while the global-reference variance value is 1.139, which is 24 percent greater.

This difference becomes even more striking when the service discipline is PS and one uses the sojourn time (instead of using the waiting time). In this case the local-reference variance measure is zero while the global-reference variance measure is 0.889.

Due to the lack of space we only state this property in this paper. A formal treatment of this subject as well as a proof that for the discrimination function $D$ of RAQFM the local-reference variance equals the global-reference variance, while for other functions it does not hold, are provided in a forthcoming work Raz, Levy, and Avi-Itzhak (2005).

## 8 Numerical Results

In Raz, Levy, and Avi-Itzhak (2004d), numerical results for the M/M/1 queue were brought, and it was demonstrated that when the seniority differences dominate service time differences (as is the case in the M/M/1 queue), RAQFM properly ranks the policies by their seniority preferences.

Our aim in the following example is to demonstrate the sensitivity of RAQFM to service time discrepancies and show that when this factor dominates the seniority factor, RAQFM reacts properly.

To this end we consider a case where the arrivals remain Poisson while the variability of service times increases drastically. This is achieved by a bi-valued service time whose values are $s = 0.1$ with probability $p$ and $s' = 10$ with probability $1 - p$. The value of $p$ is selected to be $p = 90/99 = 0.9009$ so as to have mean service time of 1, identical to the previous numerical example. The variance of this service time is $ps^2 + (1-p)s'^2 = 9.1$, in comparison to the variance of the M/M/1 case which was $1/\mu^2 = 1$. This system is analyzed via a simulation program, which was run on each evaluated point for at least $10^6$ customers. Figure 2 depicts $F_{D^2}$ as a function of $\rho$ for the FCFS, LCFS and P-LCFS cases.

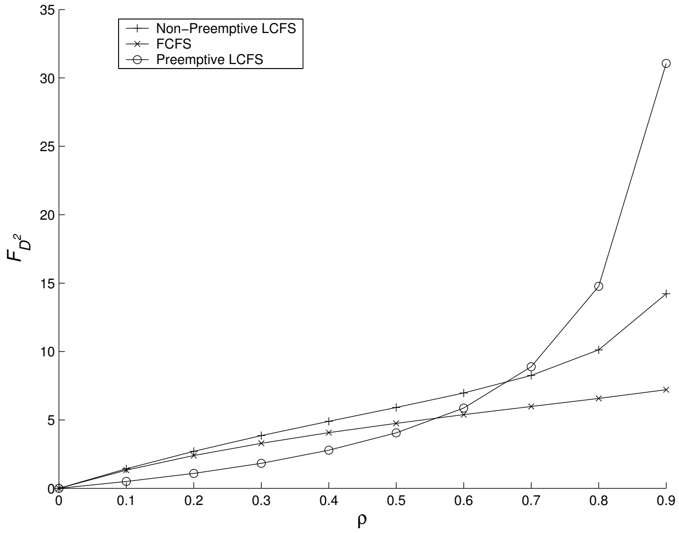


Figure 2: System Unfairness For Highly Variable Service Times

The figure demonstrates that over the range $\rho = (0, 0.55)$ RAQFM ranks P-LCFS as the most fair among the 3 policies, in contrast to its ranking in the M/M/1 case. As such, it concurs, in this case, with the ranking of the slow-down fairness approach (Wierman and Harchol-Balter (2003)) and is in contrast with the order fairness approach (Avi-Itzhak and Levy (2004)).

To understand this, note that due to the large variability of service times, large discriminations (and unfairness values) are formed when a large job is served and many small jobs queue behind it. In such cases preemption of large jobs from service can alleviate this problem. P-LCFS achieves this since it tends to preempt the large jobs with high probability. Thus, in the case of high variability service times, where service time differences dominate seniority differences, P-LCFS can be more fair than FCFS due to giving preferential treatment to short jobs over long jobs, despite its preferential service to less-senior jobs over more-senior jobs.

In summary, the example demonstrated how RAQFM accounts for the tradeoffs between seniority differences and service times differences.

It should be noted that the lower unfairness of P-LCFS does not hold for the whole range of utilizations. For high load situations P-LCFS becomes again the most unfair policy. This may possibly be attributed to the fact that at high loads the queue size tends to be large and thus the magnitude of order discrepancies may increase sharply (similarly to the waiting time variance).

## 9 Concluding Remarks

This work aimed at evaluating the RAQFM queueing fairness measure. We recognized that both *seniority* and *service requirements* must play significant roles in scheduling decisions, and showed that RAQFM accounts for both quantities. RAQFM is appropriate for measuring individual job discrimination under specific sample paths, as well as unfairness of scenarios and unfairness of systems and service policies. Further, the measure allows the use of common queueing theory techniques for evaluating the system unfairness.

We examined the sensitivity of RAQFM to seniority and service requirement and showed that in special “simple-to-understand” cases it reacts to these parameters in a proper and intuitive way. We showed that the PS service policy is uniquely absolutely fair. We further provided bounds on the measure (individual discrimination) that can be used as a scale of reference for the measure. In addition to these properties, RAQFM possesses the “locality of reference” property.

We showed that RAQFM yields to analysis of Markovian systems and demonstrated the method on the FCFS M/M/1 system.

Good understanding of fairness and proper quantification of it will allow researchers and practitioners to quantitatively account for fairness, in addition to the traditional measure of efficiency, in designing and evaluating queueing systems and scheduling policies. A comparison of the various measures of fairness in queues is important in order to better understand the subject as well as the situations in which each of the measures should be applied. Such comparison is provided in Avi-Itzhak, Levy, and Raz (2004).

## References

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

Avi-Itzhak, B., Levy, H., and Raz, D., 2004. Quantifying fairness in queueing systems: Principles and applications. Technical Report RRR-26-2004, RUTCOR, Rutgers University. URL `http://rutcor.rutgers.edu/pub/rrr/reports2004/26_2004.pdf`. Submitted.

Bansal, N. and Harchol-Balter, M., 2001. Analysis of SRPT scheduling: investigating unfairness. In *Proceedings of ACM Sigmetrics 2001 Conference on Measurement and Modeling of Computer Systems*, pages 279–290.

Bender, M., Chakrabarti, S., and Muthukrishnan, S., 1998. Flow and stretch metrics for scheduling continuous job streams. In *Proceedings of the 9th Annual ACMSIAM Symposium on Discrete Algorithms*, pages 270–279, San Francisco, CA.

Brosh, E., Levy, H., and Avi-Itzhak, B., 2004. The effect of service time variability on job scheduling fairness. Submitted for publication. URL `http://www.cs.tau.ac.il/~hanoch/Papers/Brosh_Levy_AviItzhak_2004.pdf`.

Coffman, Jr., E. G., Muntz, R. R., and Trotter, H., 1970. Waiting time distribution for processor-sharing systems. *J. ACM*, 17:123–130.

Demers, A., Keshav, S., and Shenker, S., 1989. Analysis and simulation of a fair queueing algorithm. In *Symposium proceedings on Communications architectures & protocols*, pages 1–12, Austin, Texas, USA.

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

Friedman, E. J. and Henderson, S. G., 2003. Fairness and efficiency in web server protocols. In *Proceedings of ACM Sigmetrics 2003 Conference on Measurement and Modeling of Computer Systems*, pages 229–237, San Diego, CA.

Golestani, S. J., 1994. A self-clocked fair queueing scheme for broadband application. In *Proc. IEEE INFOCOM*, pages 636–646, Toronto, Canada.

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

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

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

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

Keshav, S., 1997. *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network*. Addison Wesley Professional, Reading, MA.

Kleinrock, L., 1964. Analysis of a time-shared processor. *Nav. Res. Log. Quarterly*, 11: 59–73.

Kleinrock, L., 1967. Time-shared systems: A theoretical treatment. *J. ACM*, 14:242–261.

Larson, R. C., 1987. Perspective on queues: Social justice and the psychology of queueing. *Operations Research*, 35:895–905.

Mann, I., 1969. Queue culture: The waiting line as a social system. *Am. J. Sociol.*, 75: 340–354.

Palm, C., 1953. Methods of judging the annoyance caused by congestion. *Tele. (English Ed.)*, 2:1–20.

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

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

Raz, D., , Levy, H., and Avi-Itzhak, B., 2004a. RAQFM: A resource allocation queueing fairness measure. Technical Report RRR-32-2004, RUTCOR, Rutgers University. URL http://rutcor.rutgers.edu/pub/rrr/reports2004/32_2004.ps.

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

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

Raz, D., Levy, H., and Avi-Itzhak, B., 2004d. 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. Also appears in *Performance Evaluation Review*, 32(1):130-141.

Raz, D., Levy, H., and Avi-Itzhak, B., 2005. Locality of reference and the use of sojourn time variance for queue fairness. Forthcoming.

Rothkopf, M. H. and Rech, P., 1987. Perspectives on queues: Combining queues is not always beneficial. *Operations Research*, 35:906–909.

Wang, Y. T. and Morris, R. J. T., 1985. Load sharing in distributed systems. *IEEE Trans. on computers*, C-34(3):204–217.

Whitt, W., 1984. The amount of overtaking in a network of queues. *Networks*, 14(3): 411–426.

Wierman, A. and Harchol-Balter, M., 2003. 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.

Zhou, Y. and Sethu, H., 2002. On the relationship between absolute and relative fairness bounds. *IEEE Communication Letters*, 6(1):37–39.