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 practical scenarios. 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 propose RAQFM, a Resource Allocation Queueing Fairness Measure. The measure 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 present the methodology of RAQFM and the principles on which it is based. We discuss its properties, emphasizing the way they appeal to one’s intuition. We provide a methodology by which a wide variety of queueing systems can be analyzed to derive their fairness value.

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

Queueing systems have been used 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 queues used in all these real life situations? Perhaps the major reason for using a 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). That work uses an experimental psychology approach to study the reaction of humans to waiting in queues and to various queueing and scheduling policies. The 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 earlier by quite a number of authors. Some of these are 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 understood, and agreed upon measures do not exist. Thus, the fairness of real applications cannot be evaluated and systems cannot be compared to each other. Some published research exceptions are Avi-Itzhak and Levy (2004), Bender, Chakrabarti, and Muthukrishnan (1998), Bansal and Harchol-Balter (2001), and Wierman and Harchol-Balter (2003). 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 (e.g. Demers, Keshav, and Shenker (1990), Greenberg and Madras (1992), Parekh and Gallager (1993, 1994)); however that work does not deal with fairness to jobs, (rather it deals with fairness to streams, which fits communications systems), which is the subject of this work, and thus is outside the scope of this paper. Section 1.1 provides a discussion of these and other fairness measures proposed, and how they apply.

In light of the importance of fairness to queue-based applications, the objective of this paper is to propose a new methodology and a metric that can be applied to queueing systems and scheduling policies for evaluating their level of fairness. To devise this method we first ask what are the basic physical properties playing a role in queue fairness. We start by observing that the behavior of a queueing system is governed by two major physical factors, job seniority and customer 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). To this end, a complete fairness measure should account for both. 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 order-of-service based measure (Avi-Itzhak and Levy (2004)) focuses on the issue of seniority, with 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 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 the case of an equal service times system. In contrast, a measure that accounts only for service time differences, such as the criterion developed in Wierman and Harchol-Balter (2003), implies that Preemptive LCFS is always a fair policy, while FCFS is always unfair. The objective of this work is to propose a measure that will account both for seniority differences and for service time differences, and be convenient for analysts to work with. To achieve this, our approach is to focus on the server resources and examine how fairly they are allocated to the jobs. The approach is thus called RAQFM, a Resource Allocation Fairness Measure. 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 2. RAQFM is constructed by accounting for the individual discriminations attributed to each job in the system and summarizing them. This approach allows one to use RAQFM for several purposes: measuring individual job discrimination under a specific sample path, unfairness of scenarios, and unfairness of systems and service policies. This allows practitioners and customers to get a feel of the fairness they encounter in the system.

Having devised RAQFM we then (Section 3) turn to investigate its properties. It is important to examine such properties, especially in the context of how the measure agrees with one’s common sense intuition. 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, can assist in examining the “sanity” of the measure, developing a sense of scale to its values and building confidence in it. The first case for which we examine RAQFM is where 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. Fourth, we derive bounds on the discrimination one may be subject to; this is important in order to know the range of values of RAQFM, which can serve as a scale of reference.

Having derived the properties of RAQFM, we then (Section 4) turn to analyze it. We provide a methodology for the analysis of a Markovian type queue under steady state and carry out this analysis in detail for the single-server FCFS system. We briefly describe how to carry it out also for the single server system under the following service policies: 1) Non-Preemptive LCFS, 2) Preemptive LCFS, and 3) Random Order of Service (ROS).

Lastly (Section 5) we provide numerical results that demonstrate the unfairness of a variety of service disciplines as derived by applying RAQFM to them. Concluding remarks are given in Section 6.

1.1 Other Fairness Measures

One area where several fairness measures were proposed is flow control. The best known notion in this area is that of Max-Min Fairness (Starting with Jaffe (1981) and used by many afterwards). This notion deals with allocation of rates to customers, and it states that a rate allocation \( r = \{ r_p, p \in P \} \) (Where \( P \) is the set of sessions) is max-min fair if it is feasible, and for each \( p \in P \), \( r_p \) cannot be increased while maintaining feasibility without decreasing \( r_{p'} \) for some session \( p' \) where \( r_{p'} \leq r_p \) (Bertsekas and Gallager (1987, p. 526)). When this measure is applied to a single bottleneck node with capacity \( C \) and \( N \) connections, the only fair assignment is \( r_p = C/N, p \in P \), which is clearly achieved by PS (or indeed Round-Robin, see Hahne (1991)). Clearly, this measure applies to the assignment of rates, rather than the individual experience. This also applies to other fairness measures proposed in the area such as Proportional Fairness (Kelly (1997)), \((p, \alpha)\)-Proportional Fairness (Mo and Walrand (2000)), and lately Balanced Fairness (Bonald and Proutière (2004)).

As mentioned earlier, another close 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. When either of these measures is used for evaluating job fairness, the following emerges:

1. If the job size is 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)). Since the measures do not differentiate between different non-preemptive policies, it is inadequate for our needs.

2. Even if job size is 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 infinitely unfair, as well as Shortest Job First (SJB), Longest Job First (LJB) and 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 \( \text{PS} \), namely no job completes later under \( p \) than under \( \text{PS} \), on any sample path. This criterion is similar to AFB in that it compares the protocol against \( \text{PS} \), and it considers the worst case scenario. 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.
2 Introducing RAQFM in a Single Server System

2.1 Model and Notation

Consider a queueing system with one server. The system is subject to the arrival of a stream of customers, $C_1, C_2, \ldots$, 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 $\{a_i\}_{i=1,2,\ldots,L}$ is called an arrival pattern. A specific series of values $\{a_i, s_i\}_{i=1,2,\ldots,L}$ is called an arrival and service pattern. A specific series of values $\{a_i, s_i, d_i\}_{i=1,2,\ldots,L}$ 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.

2.2 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 momentary warranted service of $C_i$ at epoch $t$.

Summing this for $C_i$ yields $R_i \overset{\text{def}}{=} \int_{a_i}^{d_i} dt / N(t)$, 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.

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

(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 momentary discrimination of $C_i$ at epoch $t$.

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

(2)
and then the overall discrimination of $C_i$ is:

$$D_i = \int_{a_i}^{d_i} \delta_i(t) dt.$$  (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 with 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 (2004a).

2.3 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 the values $D_i$, or a function of the distribution of $D$, where $D$ is a random variable denoting the discrimination of an arbitrary customer when the system is in steady state.

Since fairness inherently deals with differences in treatment of customers a natural choice is the variance of customer discrimination. Since $E[D] = 0$, this equals the second moment and we denote this measure $F_{D^2}$. Other optional measures are the mean of distances $E[|D|]$ (denoted $F_{|D|}$) and the mean negative discrimination $-E[D|D < 0]$ (denoted $F_{D<0}$). Throughout this paper, the term “unfairness” refers to $F_{D^2}$ since the paper focuses on this measure. In some instances we also mention $F_{|D|}$. $F_{D<0}$ seems to be hard to analyze and is beyond the scope of this paper.

Note that in some cases $F_{D<0}$ might not be meaningful. For example suppose $D_1 = D_2 = \cdots = D_{19} = 1$ and $D_{20} = -19$. Then $F_{D<0} = 19$. But the average negative discrimination, assuming positive discrimination does not count, is $19/20 = 0.95$. What is more meaningful is $E[\min(D, 0)]$. However, $E[\min(D, 0)] = E[\max(D, 0)] = E[|D|]/2 = F_{|D|}/2$. 
2.4 Unfairness of a Scenario

To evaluate the unfairness of a scenario one can run (emulate) the service discipline used in the system, and compute the set of individual discriminations \( D_i, i = 1, \ldots, L \). The scenario unfairness can then be obtained via the statistical second moment (which equals the variance in our case) of these discriminations, that is

\[
\frac{1}{L} \sum_{i=1}^{L} (D_i)^2.
\]

(4)

3 Properties of RAQFM

3.1 Reaction to Differences in Seniority

In this section we show that 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" senior customers. 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.

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.
3.1.1 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 3.1 (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 unfairness (measured either as \( F_{D2} \) or as \( F_{D1} \)) of the seniority preserving schedule (a) is smaller than the unfairness of the seniority violating schedule (b), for every arrival and service pattern.

Proof. We prove the theorem for \( F_{D2} \). The proof for \( F_{D1} \) is similar, and is not shown for conciseness.

Let \( D^a_i \) denote the discrimination of \( C_i \) under scenario (a) and \( D^b_i \) the discrimination under scenario (b). Let \( F_{D2}^a, F_{D2}^b \) 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 \( \hat{F} \) denote the total unfairness to all customers other than \( C_j \) and \( C_k \), and let \( \tilde{F} \) denote the total unfairness to \( C_j \) and \( C_k \). Then \( F_{D2} = \hat{F} + \tilde{F} \).

Define \( \Delta F_{D2}, \Delta \hat{F}, \) 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. We need to prove that if \( s_j = s_k \) then \( \Delta F_{D^2} \geq 0 \), where

\[
\Delta F_{D^2} = \Delta F + \Delta \hat{F} = \frac{1}{L} \left[ (D_j^b)^2 + (D_k^b)^2 - (D_j^a)^2 - (D_k^a)^2 \right] + \frac{1}{L} \left[ \sum_{i \neq j,k} (D_i^b)^2 - \sum_{i \neq j,k} (D_i^a)^2 \right].
\] (5)

We assume now that \( s_j = \tau_1 \) and \( s_k = \tau_2 \), and denote \( \tau_{\text{min}} = \min(\tau_1, \tau_2) \) and \( \tau_{\text{max}} = \max(\tau_1, \tau_2) \). 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. \( E_4 = E_3 + \tau_{\text{min}} \)
5. \( E_5 = E_3 + \tau_{\text{max}} \)
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 in interval \( i \) (of any customer), \( R_{(i)}^a \) and \( R_{(i)}^b \) in scenario (a) and (b) respectively, is the same except for \( i = 4 \). Namely \( R_{(i)}^a = R_{(i)}^b \), \( i = 1, 2, 3, 5 \) always and \( R_{(4)}^a \neq R_{(4)}^b \) for \( \tau_1 \neq \tau_2 \).

Suppose now that \( \tau_1 = \tau_2 \). 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 \( \Delta \hat{F} = 0 \) and, from Eq.(5) we have

\[
\Delta F_{D^2} = \Delta \hat{F} = \frac{1}{L} \left[ (D_j^b)^2 + (D_k^b)^2 - (D_j^a)^2 - (D_k^a)^2 \right]
\]

\[
= \frac{1}{L} \left[ (\tau_1 - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a + R_{(5)}^a))^2 + (\tau_2 - (R_{(2)}^a + R_{(3)}^a))^2 \right] - (\tau_1 - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a))^2 - (\tau_2 - (R_{(2)}^a + R_{(3)}^a + R_{(5)}^a))^2 \right] = \frac{2}{L} R_{(1)} R_{(5)} > 0. \] (6)
Note that if $a_j = a_k$ (the two customers arrive concurrently), $\Delta F_{D2} = 0$, as expected.

**Theorem 3.2 (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 unfairness (measured either as $F_{D2}$ or as $F_{|D|}$) of the seniority preserving schedule (a) is smaller than the unfairness 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.

Define $\Phi$ to be the class of non-preemptive, non-divisible service policies (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.

**Theorem 3.3 (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.

**Proof.** Assume for the contradiction that there exists 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 adjacenty served and are not served according to their order of arrival. Interchange the order of service between these two customers (which is certainly possible). According to Theorem 3.1, 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 $\Phi$.

### 3.1.2 Stochastically Identical Service Requirement

**Theorem 3.4 (Preference of Seniority Between Adjacently Served Customers with Stochastically Equal Service Requirement).** Let $a_1, a_2, \ldots, a_L$ be an arbitrary
(deterministic) arrival pattern where \( a_j < a_k \). Let \( \{ s_i \}, i = 1, 2, \ldots, L, i \neq j, k \) be an arbitrary (deterministic) set of service times. Let \( S_j \) and \( S_k \), the service requirements of \( C_j \) and \( C_k \), respectively, be i.i.d. random variables whose values are \( s_j \) and \( s_k \). Then interchanging the order of service of \( C_j \) and \( C_k \), whenever their services are adjacent, with \( C_j \)'s being first, will result an increase in the expected overall unfairness (where expectation is taken over the joint distribution of \( S_j \) and \( S_k \)).

**Proof.** As in Theorem 3.1, the proof is given for \( F_{D2} \) and a similar proof can be given for \( F_{|D|} \).

For \( \tau_1 < \tau_2 \), \( \tau_1, \tau_2 > 0 \), consider scenario (a) of Theorem 3.1 for the case \( S_j = \tau_1 < S_k = \tau_2 \). For this case we rewrite Eq.\((5)\) as follows

\[
\Delta(F_{D2}|S_j = \tau_1, S_k = \tau_2) = \Delta(\hat{F}|S_j = \tau_1, S_k = \tau_2) + \Delta(\hat{F}|S_j = \tau_1, S_k = \tau_2). \tag{7}
\]

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

\[
\Delta(F_{D2}|S_j = \tau_2, S_k = \tau_1) = \Delta(\hat{F}|S_j = \tau_2, S_k = \tau_1) + \Delta(\hat{F}|S_j = \tau_2, S_k = \tau_1). \tag{8}
\]

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 scenario (b) with \( S_j = \tau_2, S_k = \tau_1 \). Therefore it is sufficient to show that \( \Delta(F_{D2}|S_j = \tau_1, S_k = \tau_2) + \Delta(F_{D2}|S_j = \tau_2, S_k = \tau_1) > 0 \) for the theorem to be true (from Theorem 3.1 we already have \( \Delta(F_{D2}|S_j = S_k) > 0 \)).

Following the properties shown in Theorem 3.1 we have

\[
\Delta(\hat{F}|S_j = \tau_1, S_k = \tau_2) + \Delta(\hat{F}|S_j = \tau_2, S_k = \tau_1) \\
= \frac{1}{L} \left[ \left( \tau_1 - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a + R_{(4)}^b + R_{(5)}^a) \right)^2 + \left( \tau_2 - (R_{(2)}^a + R_{(3)}^a + R_{(4)}^b) \right)^2 \\
- \left( \tau_1 - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a) \right)^2 - \left( \tau_2 - (R_{(2)}^a + R_{(3)}^a + R_{(4)}^a + R_{(5)}^a) \right)^2 \\
+ \left( \tau_2 - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a + R_{(4)}^a + R_{(5)}^a) \right)^2 + \left( \tau_1 - (R_{(2)}^a + R_{(3)}^a) \right)^2 \\
- \left( \tau_2 - (R_{(1)}^a + R_{(2)}^a + R_{(3)}^a + R_{(4)}^a) \right)^2 - \left( \tau_1 - (R_{(2)}^a + R_{(3)}^a + R_{(4)}^b + R_{(5)}^a) \right)^2 \right] \\
= \frac{2}{L} R_{(1)}(R_{(4)} + 2R_{(5)}) > 0. \tag{9}
\]
where $R^a_{(4)}$ and $R^b_{(4)}$, each appearing several times 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 $\tau_1 < \tau_2$. To clarify the relation between the four lines of expressions appearing in Eq.(9) 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 $S_j = \tau_2, S_k = \tau_1$ ($\tau_1 < \tau_2$).

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 $\Delta(F_{D2}|S_j = \tau_1, S_k = \tau_2) + \Delta(F_{D2}|S_j = \tau_2, S_k = \tau_1) > 0$.

**Corollary 3.1.** Consider a single server system with an arbitrary arrival process. Let $\{S_i\}, i = 1, 2, \ldots, L$, the service requirements of $\{C_i\}, i = 1, 2, \ldots, L$, be mutually independent random variables and let $S_j$ and $S_k$ be identically distributed for some $1 \leq j \leq k \leq L$. Then interchanging the order of service of $C_j$ and $C_k$, whenever their services are adjacent, with $C_j$’s being first (if the interchange is possible), will result in an increase in the expected overall unfairness.

**Proof.** This results from the the same argument employed in the proof of Theorem 3.4, where the equal likelihood of each pair of scenarios follows from the independence of the service times.

**Theorem 3.5 (Expected Fairness 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 expected unfairness in $\Phi$ and LCFS is the one with the highest expected unfairness.

**Proof.** The proof follows immediately from Corollary 3.1 and using an argument similar to the one used in Theorem 3.3.

Assume, for the contradiction, that there exists an arrival pattern, a service requirement distribution and a service policy $\phi \in \Phi$, $\phi \neq FCFS$, which is the service policy with the lowest expected unfairness in $\Phi$ for this arrival pattern and for this service requirement distribution. Then the order of service created by $\phi$ for this arrival pattern and for this service requirement distribution is different from the order of service created by FCFS, otherwise $\phi$ is indistinguishable from FCFS.
Given this arrival pattern, this service requirement distribution, 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. Interchange the order of service between these two customers. According to Corollary 3.1, the result of this interchange is a decrease in the expected overall unfairness. Thus the resulting order of service is more fair than $\phi$, in contradiction to $\phi$ having the lowest expected unfairness for this arrival pattern and for this service requirement distribution.

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

### 3.2 Reaction to Differences in Service Requirement

In this section we show that RAQFM reacts well to service requirement differences.

**Theorem 3.6 (Preference of Shorter Service Time, For Simultaneously Arriving Customers.).** Let $C_i$, $i = 1, \ldots, N$ 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 $a + \sum_{i=1}^{N} s_i$. 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

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

(10)

The unfairness of the scenario is

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

(11)

To evaluate Eq.(11) we first evaluate the terms involving $s_n^2$. These yield

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

(12)
Next consider the terms in the sum involving $s_n s_k$, $n > k$. These yield

$$-2 s_n (N - n) + \frac{s_k}{N - n + 1 (N - k + 1)} + \sum_{i=n+1}^{N} 2 \frac{s_n}{N - n + 1} \frac{s_k}{N - k + 1} = 0. \quad (13)$$

To summarize, the unfairness of the scenario, namely Eq.(11), reduces to

$$\frac{1}{N} \sum_{n=1}^{N} \frac{N - n}{N - n + 1} s_n^2. \quad (14)$$

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. \hfill \square

**Corollary 3.2.** 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).

The proof follows immediately from Theorem 3.6.

**Remark 3.1.** The advantage of serving a shorter service time customer $C_i$ ahead of a longer service time customer $C_j$, as in Theorem 3.6, 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

$$\{(a_i, s_i)\}_{i=1...5} = \{(0, 3), (1, 1), (1, 2), (3, 1), (6, 1000)\} \quad (15)$$

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.

### 3.3 Absolute Fairness of PS

**Theorem 3.7 (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 3.2 (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 3.7. First, PS has zero unfairness from the simple fact that for PS
\[ s_i(t) = N(t) \] for every epoch \( t \) and for every customer in the system at that epoch. Thus,
\[ \delta_i(t) = s_i(t) - 1/N(t) = 0 \Rightarrow D_i = \int_{a_i}^{d_i} \delta_i(t)dt = 0 \Rightarrow F_{D^2} = E[D^2] = 0, \] (16)
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 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 exists 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
\[ D_k = s_k - \int_{a_k}^{d_k} dt/N(t) = 0, \] (17)
where the first equality is the definition of discrimination (Eq.(1)) and the second equality was shown in Eq.(16), thus contradicting the assumption.

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,
\[ D_k' = s_k - \int_{a_k}^{d_k'} dt/N(t) = s_k - \int_{a_k}^{d_k} dt/N(t) \]

\[ = s_k - \left( \int_{a_k}^{d_k} dt/N(t) - \int_{d_k}^{d_k'} dt/N(t) \right) = \int_{d_k}^{d_k'} dt/N(t) > 0, \quad (18) \]

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,

\[ D_k' = s_k - \int_{a_k}^{d_k'} dt/N'(t) = s_k - \left( \int_{a_k}^{d_k} dt/N(t) + \int_{d_k}^{d_k'} dt/N'(t) \right) \]

\[ = s_k - \left( \int_{a_k}^{d_k} dt/N(t) + \int_{d_k}^{d_k'} dt/N'(t) \right) = - \int_{d_k}^{d_k'} dt/N'(t) < 0, \quad (19) \]

going again contradicting the assumption. \( \square \)

**Corollary 3.3 (Absolute Fairness of PS).** PS (and PS imitators) are the most fair scheduling policies.

**Proof.** For any policy \( E[D^2] \geq 0 \). For PS and PS imitators, and only for them, \( E[D^2] = 0 \) (from Theorem 3.7). \( \square \)

### 3.4 Bounds of RAQFM

In this section we derive bounds on the discrimination and unfairness measured by RAQFM.

#### 3.4.1 Bounds On Individual Discrimination

**Theorem 3.8 (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)

\[ D_i = s_i - \int_{a_i}^{d_i} dt/N(t) dt < s_i, \quad 1 \leq N(t) < \infty, s_i > 0. \quad (20) \]
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)

\[
D_i = s_i - \int_{T_S^i} dt/N(t) - \int_{T_W^i} dt/N(t) \geq s_i - \int_{T_S^i} dt - \int_{T_W^i} dt/2 = s_i - \|T_S^i\| - \|T_W^i\|/2, \tag{21}
\]

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

Note that \(\|T_S^i\| + \|T_W^i\| = d_i - a_i\), i.e. the sum of the lengths of the intervals is fixed. 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 thus \(\|T_S^i\| \leq s_i\). Thus, the minimum of Eq.(21) is achieved when \(\|T_S^i\| = s_i \Rightarrow \|T_W^i\| = d_i - a_i - s_i = W_i\). Therefore,

\[
D_i \geq s_i - s_i - W_i/2 = -W_i/2. \tag{22}
\]

To show tightness of the upper bound we let \(N(t) \to \infty, a_i \leq t \leq d_i\) in Eq.(20), 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. \(\Box\)

### 3.4.2 Bounds on System Fairness

**Theorem 3.9 (Bounds on System Unfairness).** For every scenario,

\[
0 \leq F_{|D|} < 2s_{\text{max}} \tag{23}
\]

\[
0 \leq F_D^2 < N \left(\frac{s_{\text{max}}}{2}\right)^2, \tag{24}
\]

where \(s_{\text{max}}\) 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 $s_{\text{max}}$.

The lower bounds, including their tightness, were shown in Corollary 3.3.

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:

$$-\frac{N-1}{2} < X_i < 1$$  \hspace{1cm} (25)

$$\sum_{i=1}^{N} X_i = 0.$$  \hspace{1cm} (26)

The first constraint arises from the bounds on the individual discrimination. The second constraint expresses $E[D] = 0$ (see Section 2.2).

For both $F_{|D|}$ and $F_{D^2}$, when $N \geq 3$ 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 $0 \leq X_1, X_2 \leq N - 1$ where $X_1 + X_2 = -(N - 2)$). The maximum value for $F_{|D|}$ is

$$\max F_{|D|} = \frac{1}{N} \left( \frac{(N - 1)}{2} + \frac{(N - 3)}{2} + \frac{(N - 2)}{2} \right) = 2 - 4/N < 2,$$  \hspace{1cm} (27)

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

$$\max F_{D^2} = \frac{1}{N} \left( \left( \frac{(N - 1)}{2} \right)^2 + \left( \frac{(N - 3)}{2} \right)^2 + \frac{(N - 2)^2}{2} \right) = \frac{N}{2} - 1 + 1/2N < N/2.$$  \hspace{1cm} (28)

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 \( \frac{1}{(M + N - 1)(M(N - 1)/N + (N - 1)M/N) = 2M(N - 1)/(N(M + N - 1)). \)

Taking the limiting case \( M \gg N \rightarrow \infty \), we get \( F_{|D|} \rightarrow 2. \)

4 Computing RAQFM Under Markovian Model

In this section we provide a methodology for computing RAQFM in single server Markovian type systems in steady state. We then carry out computation in detail for the single-server FCFS system (M/M/1) and describe briefly how the computation is carried out for other single server systems.

4.1 The Analysis Methodology

To facilitate the mathematical analysis of Markovian type systems, arrival and departure epochs are labeled event epochs, and time is viewed as being slotted by these event epochs. The \( l \)-th time slot, of duration \( T_l \), \( l = 1, 2, \ldots \), spans between the \( (l - 1) \)-st and the \( l \)-th event epochs. Clearly, the number of customers in the system is constant during each slot \( l \), and is denoted by \( N_l \).

We limit the analysis to systems where a service decision is made only at arrival and departure epochs. Thus, the rate of service given to each customer in the system is constant during each slot. We denote the rate at which service is given to \( C_i \) at the \( l \)-th slot by \( \sigma_{l,i} \). When a tagged customer is observed, the rate of service is denoted \( \sigma_l \). Note that the state of the system does not change during a slot (we assume that system state can only change because of either an arrival, a departure, or combined with a service decision made at such epochs).

Let the arrival and departure rates during the \( l \)-th slot be \( \lambda_l \) and \( \mu_l \) respectively. Let \( T_l, l = 1, 2, \ldots \) be the duration of the \( l \)-th slot. Then \( T_l, l = 1, 2, \ldots \) are independent random variables exponentially distributed with parameters \( \lambda_l + \mu_l \) (and first two moments \( t_l^{(1)} = 1/(\lambda_l + \mu_l), t_l^{(2)} = 2/((\lambda_l + \mu_l)^2) = 2(t_l^{(1)})^2). \)

We now move on to the methodology itself. We enumerate the steps in the analysis for reference in the following section. Consider a tagged customer \( C \). We first (step 1) define the state \( S \) that \( C \) observes in the system at any given slot. The state must include enough information for determining the number of customers in the system, the rate of service given to the tagged customer, the departure rate, the arrival rate, and for statistically predicting the following state.

21
Next (step 2), we represent the momentary discrimination in a slot in terms of the state $S$ seen at that slot, and denote it $\delta(S)$. Using Eq.(2) we have $\delta(S) = \sigma(S) - 1/N(S)$, where $\sigma(S)$ and $N(S)$ are the service rate $C$ receives at that slot, and the number of customers in the system at that slot, respectively. We also express the moments of the slot length $t^{(1)}(S)$ and $t^{(2)}(S)$ using $\mu(S)$ and $\lambda(S)$, the departure rate and the arrival rate at that slot.

Let $D(S)$ be a random variable denoting the discrimination experienced by the tagged customer, during a walk starting at state $S$, i.e., its future discrimination starting when it is at state $S$ and ending when it exits the system, and let $E$ be the state space.

Next (step 3), note that each state can change only at the end of a time slot as a result of a combined event (arrival combined with a server decisions, or departure of a customer other than $C$ combined with a service decision, or departure of $C$). For each $S_k \in E$ we enumerate the possible combined events, their probabilities and the new state caused by each. We denote by $p_{k,j}$ the transition probability from $S_k$ to $S_j$ for all $S_k, S_j \in E$, where $\sum_{\{j:S_j \in E\}} p_{k,j} = 1$. $p_{k,j}$ are computable. Note that a departure of $C$ results in an absorbing state, denoted by $S_\infty$, and conveniently $D(S_\infty) = 0$.

Assuming $C$ is in state $S_k$ in the $l$-th slot, then

$$D(S_k) = T_l \delta(S_k) + D(S_j), \text{ with probability } p_{k,j}. \tag{29}$$

*Remark 4.1.* In the equation above the sum is a sum of two independent random variables. This is so since the future discrimination, $D(S_j)$, is independent of the length of the current slot, $T_l$, and $\delta(S_k)$ is a constant.

We now let $d(S)$ and $d^{(2)}(S)$ be the first and second moments of $D(S)$. Taking expectations in Eq.(29) yields the following set of linear equations:

$$d(S_k) = t^{(1)}(S)\delta(S_k) + \sum_{\{j:S_j \in E\}} p_{k,j}d(S_j), \forall S_k \in E. \tag{30}$$

Similarity (step 4), we can take the second moments in Eq.(29). In doing so we take advantage of the fact that $\delta(S_k)$ is a constant, and that the sum represents a sum of independent random variables (see Remark 4.1) and thus the expected value of the product equals to the product of the expected values. This leads to the following set of
linear equations:

\[ d^{(2)}(S_k) = t^{(2)}(\delta(S_k))^2 + \sum_{\{j : S_j \in E\}} p_{k,j} d^{(2)}(S_j) + 2t\delta(S_k) \sum_{\{j : S_j \in E\}} p_{k,j} d(S_j), \quad \forall S_k \in E. \quad (31) \]

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

Finally (step 5), let \( P_S \) be the steady state probability that an arriving customer finds itself in state \( S \). Then

\[ F_{D^2} = \sum_{S \in \mathbb{E}} P_S d^{(2)}(S). \quad (32) \]

The probabilities \( P_S \) can usually be computed. In particular, when the arrival process is Poisson, the PASTA property (see Wolff (1982)) simplifies the computation.

Note that if state \( S \) is not possible upon arrival, then \( P_S = 0 \), and for convenience \( d(S) = d^{(2)}(S) = 0 \).

**Remark 4.2 (Conditioned Expected Discrimination and Second Moment).** For every \( S \in \mathbb{E} \), \( d(S) \) is the expected discrimination conditioned on being in state \( S \) upon arrival. The same is true for \( d^{(2)}(S) \), the conditioned second moment. These values can be used to gain insight into the behavior of discrimination and fairness for individual customers, conditioned on their arrival state.

### 4.2 Analysis of Unfairness in a FCFS M/M/1 Queue

We now carry out the analysis needed for computing unfairness in a FCFS M/M/1 queue, with arrival rate \( \lambda \) and mean service length \( 1/\mu \).

(1) Let us consider a tagged customer \( C \). At every slot let \( a \in \mathbb{N}^0 \) denote the number of customers ahead of \( C \) and \( b \in \mathbb{N}^0 \) the number of customers behind \( C \). Due to the memoryless properties of the system, the state \((a, b)\) captures all that is needed for predicting the future discrimination of \( C \).

The number of customers in the system at a slot where \( C \) observes the state \((a, b)\) is \( a + b + 1 \). The rate of service given to \( C \) at that slot is one unit if \( a = 0 \) and zero otherwise. The rates of arrival and departure are constant at \( \mu \) and \( \lambda \).

(2) The momentary discrimination at state \((a, b)\), denoted \( \delta(a, b) \), is given by:

\[ \delta(a, b) = \begin{cases} \frac{1}{a+b+1} & \text{if } a > 0, \\ 1 - \frac{1}{a+b+1} & \text{if } a = 0. \end{cases} \quad (33) \]
The moments of the slot length are,

\[ t^{(1)}(a, b) = \frac{1}{\lambda + \mu}, \quad t^{(2)}(a, b) = \frac{2}{(\lambda + \mu)^2} = 2(t^{(1)})^2, \quad a, b \geq 0, \quad (34) \]

as long as the system is non-empty (which is always true for the slots of interest, since \( C \) resides in the system). As the moments are independent of the state we can denote them \( t^{(1)} \) and \( t^{(2)} \).

(3) Let \( D(a, b) \) be a random variable denoting the discrimination of a walk starting at state \( (a, b) \), with first moment \( d(a, b) \) and second moment \( d^{(2)}(a, b) \).

Assume \( C \) is in state \( (a, b) \) at slot \( i \). At the slot end the system can encounter 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:

\[ d(a, b) = t^{(1)}(a, b) + \tilde{\lambda}d(a, b + 1) + \begin{cases} \tilde{\mu}d(a - 1, b) & a > 0, \\ 0 & a = 0. \end{cases} \quad (35) \]

(4) Similarity, the equations for \( d^{(2)}(a, b) \) are

\[ d^{(2)}(a, b) = t^{(2)}(\delta(a, b))^2 + \tilde{\lambda}d^{(2)}(a, b + 1) + 2t^{(1)}(a, b)\tilde{\lambda}d(a, b + 1) + \begin{cases} \tilde{\mu}d^{(2)}(a - 1, b) + 2t^{(1)}(a, b)\tilde{\mu}d(a - 1, b) & a > 0, \\ 0 & a = 0. \end{cases} \quad (36) \]

(5) The states possible upon arrival are \( (k, 0), k = 0, 1, \ldots \), where \( k \) is the number of customers seen on arrival. Therefore

\[ F_{D^2} = \sum_{k=0}^{\infty} p_k d^{(2)}(k, 0), \quad (37) \]

where \( p_k = (1 - \rho)\rho^k \) is the steady state probability of encountering \( k \) customers in the system.
4.3 Analysis of unfairness in a M/M/1 Queue Under Other Service Policies

In this section we state the state definitions and transition probabilities that can be used to analyze the unfairness in other service policies. For a full analysis see Raz, Levy, and Avi-Itzhak (2004b).

• Non Preemptive LCFS (NP-LCFS). Let $a \in \mathbb{N}_0$ denote the number of customers ahead of $C$ at the queue (and thus to be served after $C$). Let $b \in \mathbb{N}_0$ denote the number of customers behind $C$ at the queue (and thus to be served before $C$).

Possible events are:

1. A customer arrives into the system. The probability of this event is $\tilde{\lambda}$. If $C$ was in service ($b = 0$) $C$’s state will change to $(a + 1, b)$, otherwise to $(a, b + 1)$

2. A customer leaves the system. The probability of this event is $\tilde{\mu}$. If $C$ was in service, $C$ will leave the system, otherwise $C$’s state will change to $(a, b - 1)$.

• Preemptive LCFS (P-LCFS). The same state definition as in the non-preemptive case, however the possible events are

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

2. A customer leaves the system. The probability of this event is $\tilde{\mu}$. If $C$ was in service, $C$ will leave the system, otherwise $C$’s state will change to $(a, b - 1)$.

• Random order of service (ROS). Let $a \in \mathbb{N}_0$ denote the number of customers in the system other than $c$. Let $s \in \{0, 1\}$ be a boolean variable equaling 1 if the customer is in service and 0 if it is waiting.

Possible events are:

1. A customer arrives into the system. The probability of this event is $\tilde{\lambda}$. $C$’s state will change to $(a + 1, 0)$.

2. A customer leaves the system, and $C$ is chosen to receive service next. The probability of this event is $\tilde{\mu}/a$. $C$’s state will change to $(a - 1, 1)$.

3. A customer leaves the system, and $C$ is not chosen to receive service next. The probability of this event is $\tilde{\mu}(a - 1)/a$. $C$’s change will change to $(a - 1, 0)$. 

25
5 Numerical Results

5.1 The M/M/1 queue

Figure 2: System Unfairness For M/M/1

Figure 2 depicts $F_{D^2}$ as a function of $\rho$ for the policies studied above, under the M/M/1 queue, for $\mu = 1$. The figure demonstrates the following properties:

1. In terms of unfairness the policies may be ranked as P-LCFS > NP-LCFS > ROS > FCFS > PS. That is, Preemptive LCFS is the most unfair and Processor sharing is the most fair. This dominance is observed for all values of $\rho$, except for $\rho \to 0$.

2. At $\rho \to 0$ the unfairness measures of all policies converge, as expected, and all policies experience full fairness, $(E[D^2] = 0)$. This results from the fact that most customers find an empty system upon arrival and thus are subject to no discrimination.

3. Processor sharing is the only policy (of the policies studied) whose fairness is absolute (unfairness = 0).

4. The system unfairness seems to be monotone increasing in the system utilization $\rho$ for all policies except PS. For FCFS and ROS the increase is modest, and the values at $\rho = 0.9$ are relatively small. In fact the slopes of their curves suggest that when $\rho \to 1$ their values may converge to a finite number (in contrast, for example, to the
expected delay, which blows up at $\rho = 1$). For the LCFS policies the increase is very drastic at high loads.

It is interesting to compare these results with the policy ranking provided by the order-fairness proposed in Avi-Itzhak and Levy (2004) and the slow-down fairness approach used in Wierman and Harchol-Balter (2003). Recall that the former measure ranks FCFS as the most fair and LCFS as the most unfair, while the latter achieves a somewhat reverse ranking when it classifies LCFS preemptive as always fair and FCFS as unfair. Our results above demonstrate that in the M/M/1 case RAQFM concurs with the order-fairness as it ranks FCFS as the most fair (of the non-PS policies examined numerically) and LCFS as the most unfair. The reason for this behavior is that in the M/M/1 model the issue of seniority difference is dominant over the issue of service time differences for the compared disciplines.

5.2 The Tradeoff Between Seniority and Service Times and RAQFM’s Sensitivity to it

The prior example demonstrated that when the seniority differences dominate service time differences, RAQFM properly ranks the policies by their seniority preferences. Our aim in the next 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 next 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 3 depicts $F_{D^2}$ as a function of $\rho$ for the FCFS, LCFS and P-LCFS cases.
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 examples presented in Figures 2 and 3 demonstrate how RAQFM accounts for the tradeoffs between seniority differences and service times differences.

It should be noted that in the second example 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).
This work aimed at understanding the issue of fairness in queues and deriving a measure that can be used to quantify fairness in a variety of queueing systems. We recognized that both seniority and service requirements must play a significant role in scheduling decisions, and focused on proposing a measure, RAQFM, that accounts for both quantities. Our approach allows one to use RAQFM 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 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. Lastly, we used RAQFM to derive the system unfairness for the M/M/1 system under five fundamental service disciplines. Evaluation of special cases of the results suggest that this measure seems to agree with common intuition about fair and unfair queueing situations.

Our analysis leads to a simple ranking of the service policies examined for the M/M/1 system, independent of the load situation, according to their relative fairness (as evaluated by RAQFM). The ranking, stated from most-fair to least-fair, is PS, FCFS, ROS, NP-LCFS and P-LCFS.

Studying of queueing fairness is in its early stage and more research is needed to further understand it. Our ongoing research deals with expanding the analysis of RAQFM to a wider variety of queueing systems and service policies. In particular in Raz, Avi-Itzhak, and Levy (2004a) we studied how queue prioritization affects fairness. 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 understand the situations at which each of the measures should be applied. Such comparison is provided in Avi-Itzhak, Levy, and Raz (2004).
References


