Analysis of a Queueing Fairness Measure

Werner Sandmann

Otto–Friedrich–Universität Bamberg
Fakultät Wirtschaftsinformatik und Angewandte Informatik
Feldkirchenstr. 21, D–96045 Bamberg, Germany
werner.sandmann@wiai.uni-bamberg.de

Abstract. Scheduling policies significantly influence the performance of queueing systems, and performance measures like response and waiting times, throughput, or related properties have been comprehensively investigated in queueing and scheduling theory. The important issue of quantifying user perceived fairness has received less attention. Measuring fairness suffers from the subjective nature of fairness, and it has received growing attention only the past few years. Recently, an intuitive discrimination frequency based queueing fairness measure, which possesses important axiomatic properties, has been introduced. In this paper, we derive analytical expressions for the expected discrimination frequency in M/M/1, M/D/1 and M/GI/1 queues operating under FCFS, nonpreemptive LCFS, and SJF scheduling. Variances are evaluated by simulation and special attention is drawn to Pareto distributed service times.

1 Introduction and Motivation

Queueing models are widely used in performance evaluation of computer and communication systems. Most often, the literature has dealt with performance measures like throughput, response and waiting times. Probabilities of buffer overflows or packet losses have been considered, too. All these performance measures have in common that they are clearly defined. They do not contain subjective components in the sense that one could interpret them in different ways.

However, queueing also appears in real life scenarios, in interactive computer and communication systems and in a variety of computer networking applications, where psychological effects and individual perceptions play a crucial role in judging a system's behavior. Think for example of supermarkets, banks, airports, public offices, web servers, load balancers, or call centers. In all these and many more situations customers expect to receive a fair treatment. This is quite often even more important than response times or related measures, in particular from individual customers' points of view. Empirical evidence of the importance of fairness has been provided by the recent psychological studies in [11].

One may argue that users in computer and communication systems judge the quality of the system only through their own response time, and that there is no need to quantify fairness in such systems. However, we believe, corroborated by results in [11], that customers do care about discriminations even if they do not see each other. Hence, fairness measures are highly desirable.

Of course, fair queueing is intimately related to scheduling, which has been worked on for a long time and which is an important topic in itself [3, 4]. Fairness in queues heavily depends on the scheduling policy the system is operating under. Hence, fairness should be evaluated for different scheduling policies, and scheduling policies should be compared with regard to their fairness.

In [14] we have proposed to measure fairness by discrimination frequencies. To motivate this approach, imagine what people would think in almost all situations where discriminations occur. Being discriminated once may be accepted, but being discriminated again and again provokes the feeling of unfairness, which significantly grows with the number of experienced discriminations. What are discriminations in a queue we would call to be unfair? Other customers are served earlier although they did not arrive earlier or we have to wait for others with large service requirements, which we call *overtaking* and *large jobs*, respectively. General axiomatic properties are proven in [14], but properties for queueing systems in steady state were not yet studied.

In the present paper, we analyze the measure for M/M/1, M/D/1 and M/GI/1 systems operating under First Come First Served (FCFS), non-preemptive Last Come First Served (LCFS), and Shortest Job First (SJF). Analytical expressions are derived for the steady state expected discrimination frequency. Variance and standard deviation are evaluated via simulation, and the expected discrimination frequency is weighted by its standard deviation to account for fairness and justice in the sense of equal treatment of customers.

Section 2 describes the general framework of models our measures can be applied to. Section 3 reviews queueing fairness measures, including discrimination frequency based fairness. The analytical results are derived in section 4, and simulation results are presented in section 5. Finally, section 6 concludes the paper and gives some directions for further research.

2 General Framework

We consider general queueing systems which may consist of a single server or multiple servers and one single queue. Customers or associated jobs, denoted by $J_1, J_2, \ldots$, enter the system, are queued and served according to some scheduling policy. After service completion, jobs leave the system. For any job J_i its arrival time is denoted by a_i, where $a_i \leq a_j$ for $i < j$, i.e. jobs are numbered in increasing order of arrival. Service requirements of a job J_i are given by the job's service time denoted by s_i, and finally, the departure time of a job J_i is denoted by d_i.

Hence, relevant performance properties can be expressed in terms of a_i, s_i and d_i. In particular, $w_i = d_i - a_i - s_i$ is the waiting time of job J_i, and the job's response time is given by $r_i = w_i + s_i$. We call $t - a_i$ the *seniority of job J_i at time t*. As usual, small letters indicate that we speak of a specific sample path, and the letters capitalized are used to denote corresponding random variables. Here, a_i and s_i can be taken as realizations of random variables. Thus, in the most general case, we are dealing with G/G/c queueing systems operating under arbitrary scheduling policies.

3 Queueing Fairness Measures

A queueing fairness measure should both agree with intuition and yield to analysis. In [2, 10] desirable axiomatic properties have been introduced. The *strong service-requirement preference principle* states that for jobs J_i and J_j, arriving at the same time and residing concurrently in the system, if $s_i < s_j$ then it will be more fair to complete service of J_i ahead of J_j than vice versa. In the weak version this should hold if all jobs arrive at the same time. The *strong seniority preference principle* states that for jobs J_i and J_j, residing concurrently in the system and requiring equal service times, if $a_i < a_j$ then it will be more fair to complete service of J_i ahead of J_j than vice versa. In the weak version this should hold if all jobs have the same service time. Accordingly, there are notions of a measure adhering to and a scheduling policy following a preference principle.

3.1 Previously Proposed Queueing Fairness Measures

In [16] the slowdown $S(x) = T(x)/x$ is used as a fairness criterion, where $T(x)$ denotes the steady-state response time for a job of size x. A scheduling policy is said to be fair for given system load ρ iff $\mathrm{E}[S(x)] \leq 1/(1-\rho)$. If one takes the expected slowdown as itself instead of additionally relating it to $1/(1-\rho)$, one gets a fairness measure rather than a criterion. Obviously, slowdown relies on service requirements and does not take job seniority into account.

In [1] order fairness has been studied, where equal service times are assumed. As a major axiom monotonicity of the measure under neighbor interchange has been used, which is simply a mathematical form of the strong seniority preference principle. In steady state order fairness is equivalent to using the variance of the waiting time as an unfairness measure or with a negative sign as a fairness measure. It does not adhere to a service-requirement preference principle.

A resource allocation based queueing fairness measure (RAQFM) that aims at accounting for both seniority and service-requirements has been introduced in [13]. It relates warranted service rates to service times by subtracting the warranted service of a job from the job's service time. The measure adheres to the strong seniority preference principle and to the weak service-requirement preference principle but it does not adhere to the strong version of the latter.

3.2 Discrimination Frequency Based Fairness

The just described measures all have strengths but also some weaknesses, as mentioned in the descriptions. This motivates a measure that is intuitive, accounts for psychological effects, possesses all of the axiomatic properties, and that is analytically tractable. Here, we give a formal description of the discrimination frequency based measure followed by a brief discussion.

Definition 1. *The amount n_i of overtaking a job J_i suffers from is the number of jobs that arrived not earlier and complete service not later than J_i. That is*

$$n_i := |\{j : a_j \geq a_i \wedge d_j \leq d_i\}|.$$

To account for preemptive service disciplines, and in the non-preemptive case to account for the jobs currently in service when a new job arrives, we need to define $s'_j(t)$ as the remaining service time of a job J_j at time t. With that

Definition 2. *The amount m_i of large jobs a job J_i suffers from is the number of jobs not completely served upon arrival of J_i that have at least as much remaining service requirements and complete service not later than J_i. That is*

$$m_i := |\{j : d_i \geq d_j > a_i \wedge s'_j(a_i) \geq s_i\}|.$$

Note, that if J_i is overtaken by a job with at least the same service requirements, this affects both quantities defined above, i.e. such cases are taken as doubly unfair. This is consistent with personal feelings a customer would experience.

Definition 3. *The discrimination frequency of a job J_i is the number of discriminations J_i suffers from, that is the sum of the amount of overtaking and the amount of large jobs, i.e. $DF(i) := n_i + m_i$. The discrimination frequency of a sample path π is the sum of discrimination frequencies over all jobs.*

The measure associates higher values to such scheduling policies that are less fair. To meet the fairness measure requirements we can simply introduce a negative sign, or, alternatively, take the reciprocal (for obvious technical reasons setting to infinity if the above measure is zero). Infinite fairness then has a quite natural interpretation. The maximum fairness a job can experience is due to no discrimination at all. The most important general axiomatic properties have been proven in [14]. Here, we state them without proof as one theorem.

Theorem 1. *The discrimination frequency based queueing fairness measure adheres to the strong service-requirement preference principle, to the strong seniority preference principle and to the weak versions of both principles.*

The discrimination frequency of a sample path is monotone increasing in the sample path length. In particular, for sample path length going to infinity it goes to infinity, too. Thus, it can be used to measure fairness of a specific schedule in a specific scenario [14], but not to measure the general fairness of scheduling policies or queueing systems. System fairness is measured based on the mean discrimination frequency of an arbitrary job. When arrival and service times are given by random variables the expected discrimination frequency of a job is used. In the transient case we take the expected discrimination frequency of a job entering the system at some time t, and in steady state we take the expected discrimination frequency of an arbitrary job running through the system.

One may ask if in addition to counting the number of discriminations a degree of unfairness should be assigned to each single discrimination, like how much later a job arrived that overtakes another job or how much bigger service time a large job has. Although this seems reasonable at a first glance, results in [11] indicate, that it might be not. It has been found that people primarily care about what happens in their neighborhood. They note to be overtaken but not at what time the other job arrived. Hence, we need not distinguish the arrival times of jobs that overtake, whereas it may be reasonable to take service times into account, which is a possible but not necessary refinement.

Another question may be if it should not be taken as fair when a very short job overtakes a large job, i.e. that this is no discrimination at all. We are sure that overtaken customers will feel discriminated independent of the job size. In fact "businessmen waiting for their luggage reported feelings of injustice when they saw people that arrived after them, but had no luggage with them, leave before them" [11]. Hence, such cases should be counted as discrimination.

As for fairness not only the amount of discrimination is important but also justice in the sense of equal treatment of customers or jobs, a fairness measure should also rely on the variance or the standard deviation of the discrimination frequency. On the other hand variance or standard deviation alone should not be taken as a fairness measure, because such a measure would also assign very high fairness, when all customers experience approximately the same but large amount of discrimination. Surely, a system would not be taken as fair, when all customers are discriminated quite often. Therefore, we propose to weight the expected discrimination frequency by its variance or standard deviation.

4 Analytical Results

We derive analytical expressions for the expected discrimination frequency in M/GI/1 queues, and we also get expressions for M/M/1 as a special case. M/D/1 requires a bit more care, since all service times are equal, whereas for continuous probability distributions all services times are almost surely different.

We describe M/GI/1 queues by usual notations. Arrivals constitute a Poisson process with arrival rate $\lambda > 0$, and service times are independent and identically distributed as a random variable S according to a general probability distribution with density f_S and distribution function F_S. The server utilization is denoted by $\rho := \lambda \mathrm{E}[S]$. For comprehensive treatments of M/GI/1 queues we refer to the literature [7–9]. To avoid conflicts with other usual notations, from now on we denote the amount of overtaking by OV, and the number of large jobs by LJ.

As we are dealing with discriminations of individual jobs, it is reasonable to perform the analysis from individual jobs' perspectives. That means we inspect what jobs experience during their way through the system, instantaneously starting upon arrival and finishing upon departure. Throughout our derivations we exploit the PASTA (Poisson Arrivals See Time Averages) property[17], which basically states that in a queueing system fed by a Poissonian arrival process the distribution of jobs upon arrival of a new job is the same as in steady state.

4.1 First Come First Served (FCFS)

Obviously, under FCFS no overtaking occurs in single server queues. Thus, the frequency of this type of discrimination equals zero, and so does the mean overtaking frequency. It remains to determine the number of large jobs an individual job suffers from. Since all service times are independent and identically distributed, the probability that an arbitrary service time is greater than another arbitrary service time equals 0.5. To formally verify this, note, that due to the symmetry of the relation

$$\int_0^\infty F_S(x) f_S(x) dx = p, \quad \int_0^\infty (1 - F_S(x)) f_S(x) dx = p,$$

it immediately follows

$$p = \int_0^\infty (1 - F_S(x)) f_S(x) dx = \int_0^\infty f_S(x) dx - \int_0^\infty F_S(x) f_S(x) dx = 1 - p.$$

Hence, on average half the jobs in the queue are at least as large as the arriving job. Due to the PASTA property arriving jobs on average find the expected number of jobs in the system, in the queue, and in the server, where the latter equals the probability that the server is busy, given by the utilization ρ. The job currently in service is at least as large as the arriving job if its remaining service time S' upon arrival of the new job is greater or equal than the new job's service time. In this case the number of large jobs discriminating the job under consideration is additionally increased by one. Thus, the expected steady state number of large jobs can be expressed by

$$\mathrm{E}\left[\mathrm{LJ}\right] = \frac{\mathrm{E}\left[N_q\right]}{2} + \rho \cdot P\{S' \geq S\},$$

where N_q denotes the number of jobs in the queue. Since there is no overtaking, E[DF] = E[LJ]. The Pollaczek-Khintchine formula and Little's law yield

$$\mathrm{E}[N_q] = \frac{\lambda^2 \mathrm{E}\left[S^2\right]}{2(1-\rho)}.$$

The density $f_{S'}$ and the distribution function $F_{S'}$, respectively, of the remaining service time S' are given by

$$f_{S'}(x) = \frac{1 - F_S(x)}{\mathrm{E}\left[S\right]}, \quad x > 0,$$
$$F_{S'}(x) = P\{S' \leq x\} = \int_0^x f_{S'}(s) ds = \frac{1}{\mathrm{E}\left[S\right]} \int_0^x \left(1 - F_S(s)\right) ds, \quad x > 0.$$

Thus, we get

$$P\{S' \geq S\} = 1 - \int_0^\infty f_{S'}(x) \left(1 - F_S(x)\right) dx.$$

Combining all the above formulae yields

$$\mathrm{E}\left[\mathrm{DF}\right] = \mathrm{E}\left[\mathrm{LJ}\right] = \frac{\lambda^2 \mathrm{E}\left[S^2\right]}{4(1-\rho)} + \rho \cdot \left(1 - \int_0^\infty f_{S'}(x) \left(1 - F_S(x)\right) dx\right). \tag{1}$$

For M/M/1 we could apply the above equation, but simpler and more intuitive arguments provide the same. The remaining service time is the residual time of an exponentially distributed random variable. Due to the memoryless property it is again exponentially distributed with the same parameter. Thus, we need not make a difference between jobs in the queue and the job possibly in the server, and the expected steady state number of large jobs is half the jobs in the systems. Hence, denoting by N the number of jobs in the system,

$$\mathrm{E}\,[\mathrm{DF}] = \mathrm{E}\,[\mathrm{LJ}] = \frac{\mathrm{E}\,[N]}{2} = \frac{\rho}{2(1-\rho)}. \tag{2}$$

For M/D/1 all service times are equal and always greater than the remaining service time. Under FCFS each job suffers from all the jobs it finds in the queue upon arrival. Hence, denoting by N_q the number of jobs in the queue,

$$\mathrm{E}\,[\mathrm{DF}] = \mathrm{E}\,[\mathrm{LJ}] = \mathrm{E}\,[N_q] = \frac{\rho^2}{2(1-\rho)}. \tag{3}$$

For fixed $\rho,\ 0 < \rho < 1$ the expected discrimination frequency for M/D/1 is less than for M/M/1, or in other words, M/D/1 is more fair than M/M/1 meaning that for equal service times less discriminations are expected than for exponentially distributed service times. Note that in M/M/1 and M/D/1 the expected discrimination frequency under FCFS only depends on ρ, whereas in the general M/GI/1 case it depends on the specific service time distribution.

4.2 Last Come First Served (LCFS)

Under LCFS each job is overtaken by the jobs arriving during the waiting time in the queue. The expected steady state waiting time is the same as under FCFS:

$$\mathrm{E}\,[W] = \frac{\lambda \mathrm{E}\left[S^2\right]}{2(1-\rho)},$$

during which on average $\lambda\mathrm{E}[W]$ jobs arrive (cf. [3, 4, 7, 9]) that all overtake the arbitrary tagged job. On average half the jobs are at least as large as the tagged job and with probability ρ there may be a job in service upon arrival. Altogether

$$\mathrm{E}\,[\mathrm{OV}] = \frac{\lambda^2 \mathrm{E}\left[S^2\right]}{2(1-\rho)}, \quad \mathrm{E}\,[\mathrm{LJ}] = \frac{\lambda^2 \mathrm{E}\left[S^2\right]}{4(1-\rho)} + \rho P\{S' \geq S\},$$

$$\mathrm{E}\,[\mathrm{DF}] = \frac{3\lambda^2 \mathrm{E}\left[S^2\right]}{4(1-\rho)} + \rho \cdot \left(1 - \int_0^\infty f_{S'}(x)\,(1 - F_S(x))\,dx\right). \tag{4}$$

Note, that this equals the expected discrimination frequency for FCFS except for the factor of 3 in the first term of the sum. Hence, we have proven

Theorem 2. *The expected discrimination frequency in an M/GI/1 queue under LCFS is always greater than it is under FCFS.*

For M/M/1 we get

$$\mathrm{E}\left[\mathrm{OV}\right] = \frac{\rho^2}{1-\rho}, \quad \mathrm{E}\left[\mathrm{LJ}\right] = \frac{\rho}{2(1-\rho)}, \quad \mathrm{E}\left[\mathrm{DF}\right] = \frac{\rho(2\rho+1)}{2(1-\rho)}. \tag{5}$$

For M/D/1

$$\mathrm{E}\left[\mathrm{OV}\right] = \mathrm{E}\left[\mathrm{LJ}\right] = \frac{\rho^2}{2(1-\rho)}, \quad \mathrm{E}\left[\mathrm{DF}\right] = \frac{\rho^2}{1-\rho}. \tag{6}$$

4.3 Shortest Job First (SJF)

We start with considering large jobs. Under SJF it is only possible that a job suffers from a large job if upon arrival there is a job currently in service and the remaining service time S' is equal or greater than that of the arriving job. Thus,

$$\mathrm{E}\left[\mathrm{LJ}\right] = \rho \cdot P\{S' \geq S\} = \rho \cdot \left(1 - \int_0^\infty f_{S'}(x)\left(1 - F_S(x)\right)dx\right).$$

Similarly as for FCFS this simplifies to $\rho/2$ for M/M/1. For equal service times SJF degenerates to FCFS and the expected number of large jobs and the expected discrimination frequency for M/D/1 under SJF equal that under FCFS.

Overtaking occurs under SJF if an arriving job is smaller than some job in the queue. Compared to FCFS and LCFS an additional difficulty arises, because SJF is a job size (service time) based scheduling policy implying that the amount of overtaking depends on the job size. The conditional expected waiting time for a job of size t is given by Phipp's formula

$$\mathrm{E}\left[W(t)\right] = \frac{\lambda \mathrm{E}\left[S^2\right]}{2\left(1-\beta(t)\right)^2}, \quad \beta(t) = \int_0^t \lambda x f_S(x)dx,$$

where $\beta(t)$ is the cumulative utilization of jobs with service time t or less. During this waiting time $\lambda\mathrm{E}[W(t)]$ new jobs arrive (again cf. [3, 4, 7, 9]), and the probability for a job smaller than t is given by the distribution function $F_S(t)$. Now, we can uncondition to derive the overall expected number of overtaking:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{OV}\right] &= \lambda \int_0^\infty \mathrm{E}\left[W(t)\right] f_S(t)F_S(t)dt \\
&= \lambda \int_0^\infty \frac{\lambda \mathrm{E}\left[S^2\right] f_S(t)F_S(t)}{2\left(1-\beta(t)\right)^2}dt \\
&= \lambda \int_0^\infty \frac{\lambda \mathrm{E}\left[S^2\right] f_S(t)F_S(t)}{2\left(1-\int_0^t \lambda x f_S(x)dx\right)^2}dt \\
&= \frac{\lambda^2 \mathrm{E}\left[S^2\right]}{2} \int_0^\infty \frac{f_S(t)F_S(t)}{\left(1-\int_0^t \lambda x f_S(x)dx\right)^2}dt.
\end{aligned}$$

Combining overtaking and large jobs,

$$\begin{aligned}\mathrm{E}\left[\mathrm{DF}\right] = \quad & \frac{\lambda^2 \mathrm{E}\left[S^2\right]}{2} \int_0^\infty \frac{f_S(t) F_S(t)}{\left(1 - \int_0^t \lambda x f_S(x) dx\right)^2} dt \\ & + \rho \cdot \left(1 - \int_0^\infty f_{S'}(x) \left(1 - F_S(x)\right) dx\right) . \end{aligned} \tag{7}$$

For M/M/1 the above formula for E[OV] yields

$$\mathrm{E}\left[\mathrm{DF}\right] = \frac{\rho}{2} + \rho^2 \int_0^\infty \frac{(\mu e^{-\mu t})(1 - e^{-\mu t})}{\left(1 - \int_0^t \lambda x \mu e^{-\mu x} dx\right)^2} dt. \tag{8}$$

This expression only depends on the utilization ρ (which can be checked by an appropriate change of variables) and not on the specific values of λ and μ, which is not true for expected waiting times. Intuitively, it is clear that the quantity of interest is time-scale independent, and multiplying arrival and service rate by the same factor amounts to a change in the time scale. Hence, for M/M/1 the expected discrimination frequency under all of the three scheduling policies only depends on ρ. These results are illustrated in figure 1.

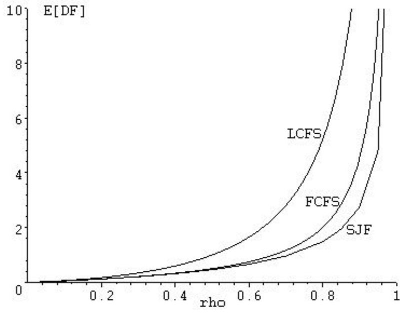

Fig. 1. Expected Discrimination Frequencies for M/M/1

5 Simulation Results

One may conclude that there is a clear ranking for fairness of FCFS, LCFS and SJF, where SJF is most fair and LCFS is least fair. But we have neither yet taken justice into account nor evaluated the expected discrimination frequency for specific non-exponentially distributed service times. Since analytical expressions are not available for the variance, we have estimated it via simulation.

We performed independent runs, where the observation period for each run was of length (=number of served jobs) 10^7 to form 99% confidence intervals with a relative half width less than 1%. Hence, the simulated values should be very accurate. LCFS yields both largest expected discrimination frequency and largest variance, and thus we can neglect it in our comparison.

5.1 Exponentially Distributed Service Times

Figure 2 shows the standard deviation and the product of the expected discrimination frequency and the standard deviation for M/M/1 in the range of $\rho > 0.7$, where this product is significantly different for FCFS and SJF.

Now, we get a different impression of the fairness and justice of FCFS and SJF in M/M/1 queues. It is clear that in case of using variance instead of standard deviation the curves for FCFS and SJF differ even more. Simulations for Erlang distributed service times yield similar pictures as above. Obviously, the justice of SJF is less than the justice of FCFS, although the expected discrimination frequency is greater under FCFS. This directly implies that in applications where equal treatment of customers or jobs is most important, FCFS may be preferred.

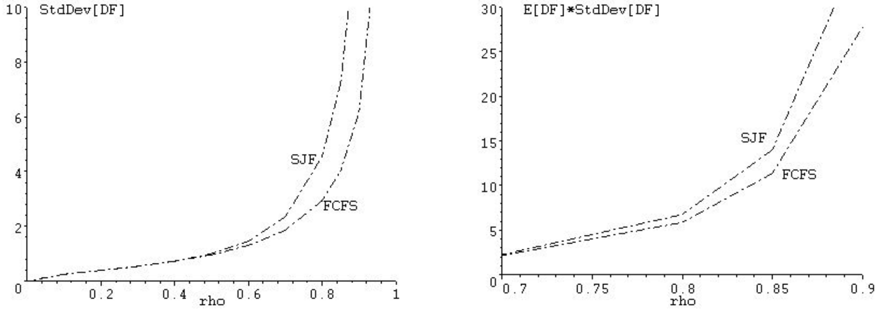

Fig. 2. Justice related measures for M/M/1

5.2 Pareto Distributed Service Times

Consider Pareto distributed service times with density and distribution function

$$f(x) = \frac{ab^a}{x^{a+1}}, \quad F(x) = 1 - \left(\frac{b}{x}\right)^a \quad a > 0, x > b.$$

The expectation only exists for $a > 1$, the second moment only for $a > 2$:

$$\mathrm{E}[X] = \frac{ab}{a-1},\ a > 1; \quad \mathrm{E}[X^2] = \frac{ab^2}{a-2},\ \mathrm{VAR}[X] = \frac{ab^2}{(a-2)(a-1)^2},\ a > 2.$$

Unfortunately, for Pareto distributed service times, evaluation of the formerly derived analytical expressions requires computations of the transcendental Lerch-Phi function [5, 6, 15]. Thus, we utilize simulation. It can be easily checked that the expected discrimination frequency in M/Pareto/1 queues does not depend on the location parameter b, but only on the scale parameter a. Therefore, we can restrict to $b = 1$, a special version of the Pareto distribution. We present results for two values of a, namely $a = 10$ and $a = 2$. Note, that in the first case, the variance of the Pareto distribution is very small, whereas in the second case, the variance even does not exist.

Figures 3–5 show the expected discrimination frequency for $a = 2$ and $a = 10$, and the justice related measures, i.e. the discrimination frequency's standard deviation and the product of the expected discrimination frequency and its standard deviation, both for $a = 2$ and $a = 10$.

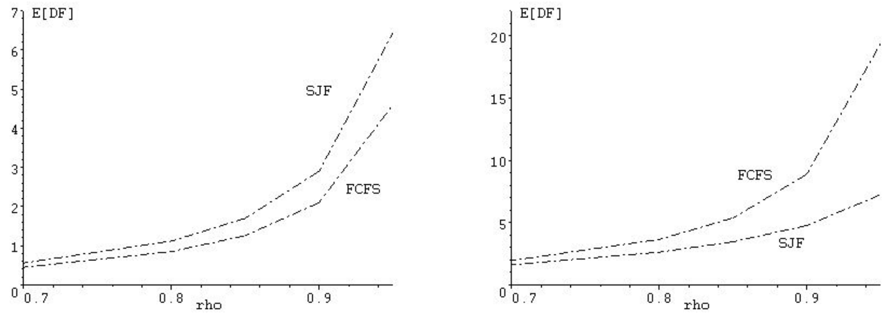

Fig. 3. Expected discrimination frequency for $a = 10$ (left) and $a = 2$ (right)

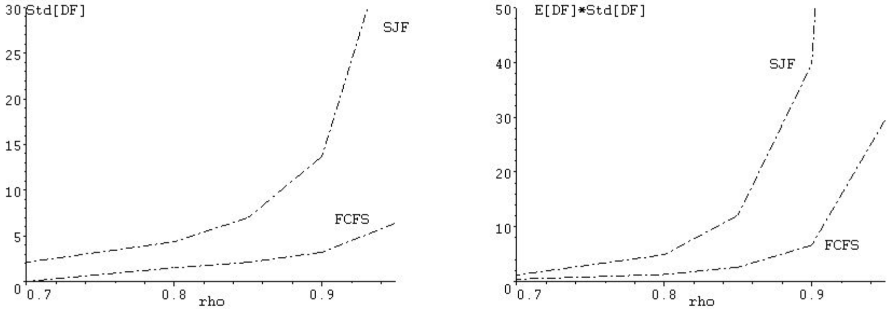

Fig. 4. Justice related measures for $a = 10$

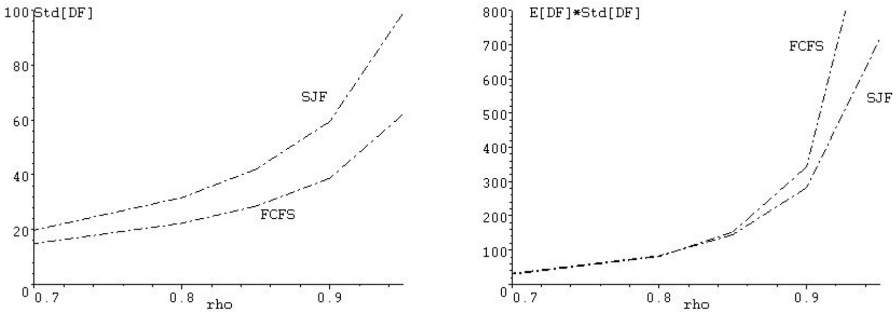

Fig. 5. Justice related measures for $a = 2$

For $a = 10$ we observe that SJF is more unfair than FCFS in terms of all the three measures, which is an important difference to the M/M/1 system. One could conclude, that FCFS should be taken as more fair than SJF for M/Pareto/1 queues, but the three measures for $a = 2$, where the variance of the Pareto distributed service time does not exist, take us to a different conclusion.

In this case, as for M/M/1 queues, SJF is more fair regarding only the expected discrimination frequency and FCFS is more fair regarding the standard deviation of the expected discrimination frequency. In opposite to M/M/1 queues, the product of the expected discrimination frequency and its standard deviation is less for SJF than it is for FCFS.

Obviously, it is not clear, which scheduling policy should be taken in general. When designing a fair system, the scheduling policy should be chosen with regard to the specific involved distributions and its parameters. One cannot generally say, that if fairness or justice is the main system design goal, then take policy x or y. This is an important difference to minimizing waiting or response times as the main goal. In this case, it is well known, that one has to take SJF.

6 Conclusion and Further Research

We have derived analytical expressions for the expected discrimination frequency in M/M/1, M/D/1 and M/GI/1 queues under FCFS, LCFS and SJF. For Pareto distributed service times the discrimination frequency has been evaluated by simulation. Simulation results for the standard deviation and for weighting the expected discrimination frequency by its standard deviation have been provided.

The results indicate, that LCFS is the least fair of the considered policies. In M/M/1 queues FCFS is less fair than SJF when only accounting for expected discrimination frequency, whereas FCFS is more fair than SJF when additionally accounting for justice. In M/Pareto/1 queues a fairness ranking of scheduling policies depends on the specific distribution parameters. Our results particularly imply that in system design the decision on how to schedule customers has to be made very carefully and should consider specific system parameters.

Some refinements may improve our fairness measure by a differentiated view on the severity of discriminations. It may be reasonable to rate a discrimination in such a way, that the first overtaking is weighted less than the second overtaking, the second overtaking weighted less than the third overtaking and so on. Besides, overtaking by very much smaller jobs may be weighted less than overtaking by jobs with nearly the same or even larger service requirements. Discriminations due to waiting for large jobs may be weighted by the job size.

Comprehensive simulation studies for non-preemptive and preemptive scheduling in G/G/c queues are part of current research. Further analytical investigations for M/GI/1 and more general models are particularly desirable, e.g. deriving expressions for the variance and the standard deviation or for the distribution of discrimination frequency. Another important topic is to extend fairness measures in general, and in particular the discrimination frequency based fairness measure, to networks of queues.

References

1. Avi-Itzhak, B., Levy, H.: Measuring fairness in queues. Advances in Applied Probability 36 (2004) 919–936
2. Avi-Itzhak, B., Levy, H., Raz, D.: Quantifying fairness in queueing systems: principles and applications. Rutcor Research Report (2004)
3. Coffman, E.G., Denning, P.J.: Operating Systems Theory. Prentice Hall (1973)
4. Conway, R.W., Maxwell, W.L., Miller, L.W.: Theory of Scheduling. Addison-Wesley (1967)
5. Erdelyi, A.: Higher Transcendental Functions. McGraw Hill (1953)
6. Gradshteyn, L.S., Ryzhik, L.M.: Table of Integrals, Series, and Products. Academic Press, 6th ed. (2000)
7. Haverkort, B.R.: Performance Evaluation of Computer Communication Systems – A Model-Based Approach. Wiley (1998)
8. Kleinrock, L.: Queueing Systems Volume I: Theory. Wiley (1975)
9. Kleinrock, L.: Queueing Systems Volume II: Computer Applications. Wiley (1976)
10. Levy, H., Avi-Itzhak, B., Raz, D.: Principles of fairness quantification in queueing systems. Proc. HET-NETs (2004) T1/1–17
11. Rafaeli, A., Kedmi, E., Vashdi, D., Barron, G.: Queues and fairness: a multiple study experimental investigation. Manuscript under review (2005) http://iew3.technion.ac.il/Home/Users/anatr/JAP-Fairness-Submission.pdf
12. Raz, D., Avi-Itzhak, B., Levy, H.: Classes, priorities and fairness in queueing systems. Rutcor Research Report (2004)
13. Raz, D., Levy, H., Avi-Itzhak, B.: A resource allocation queueing fairness measure. Performance Evaluation Review 32 (2004) 130–141
14. Sandmann, W.: A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement. Proc. NGI (2005) 106–113
15. Weisstein, E.W.: Lerch transcendent. MathWorld - A Wolfram Web Resource. http://mathworld.wolfram.com/LerchTranscendent.html
16. Wierman, A., Harchol-Balter, M.: Classifying scheduling policies with respect to unfairness in an M/GI/1. Proc. ACM SIGMETRICS (2003) 238–249
17. Wolff, R.W.: Poisson arrivals see time averages. Operations Research 30 (1982) 223–231