R U T C O R
R ESEARCH
R E P O R T

QUANTIFYING FAIRNESS IN QUEUEING SYSTEMS: PRINCIPLES, APPROACHES AND APPLICABILITY

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

RRR 25-2005, REVISION OF RRR 26-2004, AUGUST 2005

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

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

RUTCOR RESEARCH REPORT
RRR 25-2005, REVISION OF RRR 26-2004, AUGUST 2005

QUANTIFYING FAIRNESS IN QUEUEING SYSTEMS: PRINCIPLES, APPROACHES AND APPLICABILITY

Benjamin Avi-Itzhak Hanoch Levy David Raz

Abstract. In this paper we discuss fairness in queues, view it in the perspective of social justice at large and survey the recently published research work and publications dealing with the issue of *measuring* fairness of queues. The emphasis is placed on the underlying principles of the different measuring approaches, on reviewing their methodology and on examining their applicability and intuitive appeal. Some quantitative results are also presented.
The paper has three major parts (sections) and a short concluding discussion. In the first part, fairness in queues and its importance are discussed in the broader context of the prevailing conception of social justice at large. A special effort, including illustrative examples, is made to differentiate between *fairness of the queue* and *fairness at large*, which derives from favoring the more needy. The second part is dedicated to explaining and discussing the three main properties expected of a fairness measure: conformity to the general concept of social justice, granularity, and intuitive appeal and rationality. The third part reviews the fairness of the queue evaluation and measuring approaches proposed and studied in recent years. We describe the underlying principles of the different approaches, present some of their results and review them in context of the three main properties expected from a measure.
The short discussion that follows centers on future research issues.

Acknowledgements: This work was supported in part by grant 380-801 from the Israeli Ministry of Science and Technology, and by the Euro NGI network of excellence.

1 Introduction

1.1 Preface

Why are we using ordered queues? Why do they serve in many real life applications, such as banks, supermarkets, airports, computer systems, communications systems, Web services, call centers and numerous other systems?

While the major reason for the formation of queues is economic, i.e. scarcity of resources, the dominant reason for using ordered (disciplined) queues is often the strive to maintain some level of social justice, or in other words fairness in treatment of everyone involved.

In this sense, a system serving a queue of people is a microcosm social construct. Emotions and resentment may flare if unfairness is practiced, or is perceived as being practiced, while courtesy, and even camaraderie due to same experience-sharing, may result when fairness in treatment is perceived (see Rafaeli et al (2002)). Notwithstanding its fundamental role, the fairness factor was virtually neglected, or even disregarded, in the published queueing literature until quite recently. Aspects of fairness in queues were recognized and discussed, or mentioned in passing, quite early by a considerable number of authors: Palm (1953) deals with judging the annoyance caused by congestion, Mann (1969) discusses the queue as a social system and Whitt (1984) addresses overtaking in queues, to mention just three.

While almost every child, if asked, can tell you what is fair and what isn't, it is not an undemanding undertaking to have a group of people agree on a common definition of fairness, much more so when it comes to defining a quantitative measure of the level of fairness, and when the group is vast. It is not surprising, then, that extensive research aimed at developing fairness measures for queues, in contrast to the traditional "efficiency" measures of sojourn and waiting times, has been slow in coming.

Traditionally, a first-come-first-served (FCFS), or a first-in-first-out (FIFO), queue discipline is considered most fair. This probably derives from experience in queues where the total amount of service the system is able, or willing, to dispose is limited by a maximal number served or by a length of time the system is open for service, i.e. exhaustible-servers systems. In such systems (e.g. a line at a gas pump at a time of energy crisis, a line for basic foods in a refugee camp, or less dramatic, a line for tickets for a show or a sport event), which were very prevalent in the human experience, if you are not early enough in the queue, chances are you will never get the service, or product, or you may have to come again at a future time. Namely, the early bird gets the worm. Placing ahead of you a person, who arrived after you, will be regarded as grossly unfair, particularly if that person is not needier than you. Many present day queueing systems, however, are not of this type, rather, all birds get their worms, not only the early ones, and thus FIFO may not be as crucial in these systems. Fairness of exhaustible-servers queues is an important issue, deserving attention on its own, and is outside the scope of this paper, that focuses on non-exhaustive servers.

Larson (1987) in his discussion paper on the psychology of waiting recognizes the central role played by 'Social Justice', (which is another name for fairness). In the first part of his paper, dedicated to social justice in queues, he brings several anecdotal actual situations, experienced by him and others (Martin (1983), Kettelle (1986) and Lewin (1986)), that strongly support the

traditional claim of FIFO being the most socially just queue discipline. In fact he practically defines social injustice as violation of FIFO when stating "…customers may become infuriated if they experience social injustice, defined as violation of FIFO."

What would be a fair service order in a supermarket queue or in an airport waiting line? Many people would instinctively embrace Larson's view, responding that FIFO is the fairest order, that is, serving the most *senior* customer first, where *seniority* is measured in the time the customer has already spent in the line. Already Kingman (1962) pronounces this same view by calling FIFO "the fairest queue discipline". The underlying principle, or rationale, of this view can be expressed in one sentence: *the one who has been waiting longest earned the right to be served first*. But, recalling that the server is non-exhaustible, *is* FIFO undeniably the most fair discipline?

To answer this question, consider a common situation at a supermarket counter, which some readers may associate with their own personal experience: Mr. Short arrives at the supermarket counter holding only one item. In the line ahead of him he finds Mrs. Long carrying a fully loaded cart of items. Long says to Short "Excuse me, I only have one item. Would you mind if I go ahead of you?" Would it be fair to have Mr. Long served ahead of Short and Short waiting for the full processing of Mrs. Long's loaded cart? Or, would it be more fair to advance Short in the queue and serve him ahead of Long? This dilemma may cause some to "relax" their strong belief in the absolute fairness of FIFO. In fact, the dilemma brings to the discussion a new factor, that of *service requirement*. The basic intuition thus suggests that prioritizing short jobs over long jobs may also be fair, based on the underlying principle: *the one who demands the least of the server's time should be served first*. It is the trade-off between these two factors, *seniority* (prioritize Mrs. Long) and *service requirement* (prioritize Mr. Short), that creates the dilemma in this case. To demonstrate the conflict we continue our scenario in two directions: (i) Long looks at Short, smiles and says "Why don't you go ahead of me. I have arrived only a few seconds ago and it is not fair that you will wait that long while your short service will delay me very little". This is one possibility. Alternately, Long may be negative, saying (ii) "Look, I have been waiting in this line forever. If not for this lengthy wait I would have been out of here long before your arrival. You can patiently wait too". Clearly, Long weighs their seniority difference against their service requirement difference in deciding what is the fair thing to do. This tradeoff, illustrated by the "Long vs. Short" scenario, will accompany us in this paper in attempting to understand fairness in queues. (It should be noted that many supermarkets handle this conflict by allocating some of the counters exclusively to Shorts who also retain the option to select a "regular" counter.)

Figure 1

1.2 What is "Fairness of the Queue"?

Evidently, there is a need to agree upon the definition of fairness, or at least the underlying principles, or rationale, that forms its foundation. As mentioned earlier, a queueing system is a microcosm social construct and its fairness should conform to the general cultural perception of social justice in the particular society. Social justice has always been, and still is, a cardinal issue in all cultures. It is the cement holding the society together. As such it has been subject to debate by philosophers, prophets and spiritual leaders since the beginning of recorded history. In modern time, many economists and social scientists joined the ongoing debate. Since social justice perception is culture and time dependent we are interested in the modern western societies perspective. As is to be expected, there is a vast ocean of modern research and publications on this issue, mostly by philosophers, economists and social and behavioral scientists. Reviewing and interpreting this literature is much beyond the scope of this paper, and probably also beyond our ability. To readers who would like to dive into this ocean, or just wet their feet at its shores, we recommend to start with visiting the Stanford Encyclopedia of Philosophy (2005). A most, some would say *the* most, prominent and comprehensive publication on this issue is Rawls' book "A Theory of Justice" (1971, 1999)[1]. The book does not make for an easy reading, but in essence, Rawls' general conception of social justice, as summarized in a nutshell by Piccard (2005), is:

> **All social primary goods – liberty and opportunity, income and wealth, and the bases for self-respect – are to be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored.**

We are back to the traditional economists' approach of achieving social justice by appropriately dividing the "pie", except that the pie here is made of a mix of tangibles and non-tangibles, while the traditional economists' pie is wholly tangible. By Rawls conception, if all persons involved are equally non-favored (equally needy) the pie should be equally divided.

[1] Nussbaum (2001) describes Rawls as "the most distinguished moral and political philosopher of our age".

Obviously, Rawls' conception, though widely recognized, has its dissenters, as is true for practically any social issue. We will use it here as a guideline.

1.2.1 Fairness of the Queue vs. Fairness in a Queueing System

Social justice in a queueing environment, i.e. a queueing system, does not differ from social justice at large and if we accept the above conception it must also apply to queueing environments as well. We therefore need to differentiate between *fairness in a queueing system* and *fairness of a queue*, which is, roughly put, the fairness component that is attributable to the queue discipline or structure. For illustration, imagine a waiting room packed with patients. The door to the doctor's office opens and a nurse appears and asks: "Who is the sickest?" This order of service is near to longest–job–first, LJF. Still, the many, if not most, will say it is fair by the principle that those most at-risk, or those suffering the most, should be attended to first. The fairness issue is cast here in a queueing situation. Alas, it has little to do with fairness of the queue. Very few will categorize a LJF ordering as fair, given that all customers are equally needy.

Figure 2

In this discussion paper we define the *fairness/unfairness of the queue* as the fairness/unfairness that can be related to the discipline or configuration of the queue when all customers are equally needy. Customers will be assumed to be equally needy if they are only discernable by their arrival time and service requirement, and are identical in all other respects. The doctor's waiting room scenario, which was artificially constructed to make a point, is very realistic when looking at hospitals' ERs. Arriving patients are categorized into several classes of neediness (urgency, or critical level, of condition) and the classes are prioritized in accordance to their level of neediness[2]. The fairness related to the class prioritizing is determined by the nature

[2] I learned about ER prioritizing both the hard way and the easy way. The hard way was when, about a dozen years ago, I had to "visit" a local ER due to a relative minor injury requiring several stitches. It didn't take me long to realize that FIFO order was not followed. During a couple of hours wait I observed quite a number of severely

of the required service and neediness of the customers. These may vary widely from one queueing system to another and a universal measure to quantify the related fairness is not likely to be found[3]. The "fairness of the queue" is related to the order of service within each class, under the assumption that same class patients are practically equally needy.

Note that most customers would not differentiate between fairness of the queue and the fairness of the system, unless specifically guided to so do. Rafaeli et al (2005, Study III) conducted an experiment comparing perceived fairness by customers in a multi-server/multi-queue system (each server has its own queue, served in a FIFO order) to that of customers served in the same system that has, in addition to the regular queues, VIP queues (e.g. business class check-in counters in an airport). Only responses of those served in non-VIP queues were considered. Average fairness in the VIP structure was found to be significantly lower than in the structure without VIP queues, unless people knew that those in the VIP queue had paid a special fee in order to join it. That is, the queue was perceived as unfair by participants who thought others are getting a preferential treatment with no justification. However, once they learned that the preference was bought for a special fee they perceived the same system as being fair. In the first situation we are dealing with the perceived fairness of the queue. In the second situation we are dealing with the perception of the system's fairness at large. Participants perceive buying preferential treatment as fair.

In what follows, fairness is meant to stand for fairness of the queue, unless otherwise specified. The distinction between customers based on their neediness, "value" or other economic factors[4] is not considered.

1.2.2 What is the Pie?

Assume all customers are equally needy, what is then the "pie" and how can it be equally divided? Clearly the scarce resource, or the pie, is the service rendered by the servers. Consider an M/D/1 system where service requirement is the same constant for all customers. In such a system all customers seemingly receive an equal share of the server's attention, hence FIFO and LIFO are equally fair! Not so, the pie's division is not a one-time act. In an on-going process the pie must be continually divided, i.e. timely divided. Therein lies the key to the just division.

One can take either a circuitous approach or a direct one to attaining a just timely division of the resource. In the circuitous approach the customer is assumed to get the utility of the service

---

injured new patients, mostly accidents victims, being treated before me. I didn't perceive this as being unfair, in fact, I remember the guilt I felt at taking even the twenty minutes of a doctor's time, which I felt was taken away from much needier patients. The easy way was when, a couple of years ago, one of our UG seniors approached me, requesting my supervision of an independent project. I found that he was doing an internship in a local hospital, and after some inquiry, he told me that the hospital is considering expanding its ER. He ended up putting together a queueing model to help determine how many beds will be needed. I ended up learning in detail quite a lot about ER operation. Ben Avi-Itzhak.

[3] A possible "universal" way to address this is to model neediness levels by customer weights and study fairness in this framework; in this case the burden of dealing with various problems is on selecting the weights (based on the problem) while the queue fairness measure is uniform. This is a subject for future research.

[4] There is a large body of literature on the aspects of queue pricing and its relation to scheduling and customer behavior in the queue (see, e.g., a recent book by Hassin and Haviv (2002)). The relation between pricing and fairness is beyond the scope of this article and requires further study. See further discussion at the end of Section 4.

plus the disutility of the wait. Therefore in the M/D/1 case, where equal service time is given to all customers, the waiting times must also be identical, to attain *absolute* fairness. Unfairness in this case is produced by deviations from equal wait, and a "natural" measure for it is the waiting time variance. The fairest discipline must produce the smallest waiting time variance. For the M/D/1 class of disciplines that are non-preemptive (Processor Sharing is considered to be preemptive; see discussion of PS in Section 3.2) and work conserving, the smallest variance is produced by FIFO. This is also true for M/G/1, (Kingman (1962), Avi-Itzhak and Levy (2004)). In extending this approach to the M/G/1 system we note that customers receive unequal shares of the server's attention, giving rise to the Long-versus-Short dilemma. One way to solve this conflict is to assume that absolute fairness is achieved if the waiting disutility of each customer is proportional to his service utility (assuming, for simplicity, linearity of both utility functions). The unfairness measure can be derived from the variability, or the normalized variability, of the deviations from this proportionality. In both cases of the circuitous approach, the particular "equal and timely pie division" results from a pragmatic *perceived* fairness of the queue, instead of vice versa. The conformity to the conception of general social justice is an after-the-fact rationalizing of the two pragmatic fairness-of-the-queue principles used.

The direct approach to dividing the pie, assumes that the community entitled to a slice of it at any time point is made of the customers present in the system at that time. If there are N customers present, each is entitled to receive ($1/N$)-th of the servers' attention (service rate), to achieve absolute fairness. Thus the pie is equally divided at all points in time. Deviations from this division of the server's rate are unfair, and a summary measure of their variability can serve as an unfairness measure. In this approach the absolute fairness results from the just division of the pie, in contrast to the circuitous approach. Still, it remains to agree upon the definition of a deviation from the defined just division of the servers' rate.

1.3 Importance and applicability of fairness of the queue

As already mentioned, the fairness factor has long been recognized in queueing literature. Larson (1987) brings several actual situations where fairness considerations play a role in deciding the structure and discipline of service systems. Nevertheless, queueing theory has been mostly occupied with the performance metrics of waiting time. This has been the main quantity (perhaps almost the sole quantity) used in queueing theory to evaluate queueing systems (see, e.g. text books on the subject, Gross and Harris (1974), Kleinrock (1975, 1976), Hall (1991), Cooper (1981), Daigle (1992)) and is frequently being looked at via the expected delay or its variance in steady state. Under this quantity, customer satisfaction decreases with the delay experienced by the job and thus customer's objective is to minimize delay. The use of this quantity seems to be appropriate when the major performance issue associated with job queueing is indeed the delay experienced in the system. The fairness factor, though playing an important role in the design and operation of actual waiting systems, has recently become a topic of interest also to queueing theorists. Rothkopf and Rech (1987), in their paper discussing perceptions in queues, bring an impressive list of quantifiable considerations showing that combining queues may not be economically advantageous, contra to the "common" belief. At the end they concede however, that all these considerations may not have sufficient weight to overcome the unfairness perceived

by customers (as suggested by Larson (1987), based on a private communication by A. Lewin) served in a separate queues structure.

Experimental evidence of the importance of fairness in queues was recently provided in Rafaeli et. al. (2002), who studied, using an experimental psychology approach, 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 some times more important than the duration of the wait. For the case of common queue versus a separate one at each server, they found that the common queue was perceived as more fair. Probably for this reason we find separate queues mostly in systems where a common queue is physically not practical, e.g. traffic toll boots and supermarkets.

Fairness and efficiency are the major reasons for the need for disciplined queues. In queues, like in most situations of limited resources, there is a need to utilize, or share, the resources in an efficient and fair way. Thus an ordered queue is a fairness and efficiency management facility and is perceived as such by most service systems operators, particularly those subject to competition. Supermarkets, where common queues are not always practical, try to increase both fairness and efficiency by assigning some of the counters to Shorts only. The same practice is common to toll booths as well. An alternate solution is to make a common queue feasible by allocating the necessary additional resources. For example, if you arrive to Newark airport on an international flight you find that the passport control queue is common and an extra attendant is assigned, to orderly direct people to the next available server as to reduce overtaking.

1.3.1 How Applicable is Fairness of the Queue in a 'Blind Queue'

In the course of our study of fairness in queues we were asked more than once "is fairness relevant at all in a blind queue?" There are many situations where customers cannot see each other and are not informed of the state of the system and the discipline used. Call centers know from experience that some customers are impatient and are likely to renege after a relatively short wait. More patient customers will hang on for quite a while before hanging up (pun not intended). Therefore, a waiting customer is more likely to be a patient one, as compared to a new arrival. Using LIFO waiting line discipline will result in retaining more customers and increased profit. However, most customers would consider LIFO as unfair, even if informed of it ahead of time, and outrageous if it is concealed and then revealed to them somehow. In fact, in today's information age it is hard to expect such practice to remain concealed for long time. Suppose, nonetheless, that such LIFO practice can indeed be hidden. Does it make the practice fair? No. Is fairness in this case relevant? This is a question of ethics. Is cheating right if it never gets disclosed and the cheater can get away with it unscathed? The answer to this question, like the answer to the question of whether fairness is relevant in a blind queue, depends on your ethical values.

In fact, making the queue less blind might be quite important to customers. Many call centers will inform you of your place in the line and sometimes provide you with an estimate of the wait involved. This allows you to be aware that the order of service is FIFO and enables you to renege now, instead of wasting so much of your time before reneging anyway. Both are fairness considerations. Along these lines, surveys of 911 callers who were classified by the police as "low priority", and kept waiting a long time for police arrival, found that callers were not

dissatisfied with the service, provided they were told that the police are busy with higher priority calls and tasks, and were also told to expect a long delay (see Larson (1987), Chan and Tien (1981) and McEwen, Connors and Cohen (1984)). In this case, though we are dealing with fairness based on need rather than fairness of the queue, the knowledge that the system is fair strongly influences the callers' degree of satisfaction and prevents repeated calls and complaints.

1.4 Flow Related Fairness

Queueing model applications can be classified into 1) *Job-based systems*, and 2) *Flow-based systems*. In the former, the *i-th* customer, say C_i, is associated with a single job J_i arriving at epoch a_i. Of interest is therefore the performance experienced by that individual job, which is synonym to customer in this paper. In the latter, customer C_i is associated with a stream (or flow) of jobs $J_i^1, J_i^2, \ldots$ arriving at epochs $a_i^1, a_i^2, \ldots$ respectively. Of interest is the performance experienced by the whole flow. The applications associated with this latter model are communications networks applications where a customer (sometimes called source) is associated with a stream of packets that are sent through a communications device, e.g., a router.
Much work has been done and published in the context of communications networks where the concern is with *flows* traversing a communications node and in allocating the *bandwidth fairly* among the *flows.* This is in contrast to the present work that focuses on fairness to *jobs*. One of the earliest attempts to define flow-fairness is Wang and Morris (1985) where the Q-factor is defined. Later on, the research on flow fairness has flourished with the introduction of *Weighted Fair Queueing* (WFQ), which deals with the fair scheduling of packet flows. Some early papers on the subject are Demers, Keshav and Shenker (1990), Greenberg and Madras (1992), Parekh (1992), Parekh and Gallager (1993), (1994), Golestani (1994), Rexford, Greenberg and Bonomi (1996), Bennet and Zhang (1997). Many other papers have been published on this subject. A popular measure of fairness within that context is the *relative fairness bound* (Used by Golestani (1994) and others) which captures the maximum possible difference between the (normalized) service received by any two streams. As such it measures "fairness of throughput of streams"[5].
Our focus in this work is on job-based systems. In what follows, customer C_i and job J_i are synonymous and will be used interchangeably. Applications that are associated with this model are:

1. **Banks, supermarkets, public offices and the like**, in which customers physically enter queues where they wait for service and then get served.
2. **Some computer systems,** in which a customer (or a customer's computer application) submits a job to the system, and the customer gets satisfied when the service of the job is completed.
3. **Call Centers,** in which customers call into a call center to receive service, possibly wait in a virtual queue (while listening to some music)**,** until being answered by "the next available agent". Call center queueing systems are conceptually identical to physical queueing

[5] Within the context of a network, the literature deals with fair allocation of bandwidth (e.g. the *Max-Min fairness* allocation (Jaffe (1981)), *Proportional fairness* (Kelly (1997)), and *Balanced Fairness* (Bonald and Proutiere (2004)), which is orthogonal to our work.

facilities, such as banks or airlines counters, except that the queue can be blind unless the operator decides otherwise.

2 Properties expected of a fairness measure

When dealing with the introduction of a new queueing performance measure for an entity that is somewhat abstract and not very tangible, several questions should be brought up and discussed. What is the underlying principle or conception that is in the foundation of the measure? Does this principle conform to the wider, non-queueing related, approach to dealing with this entity? What is the *physical quantity,* or *performance objective* that should be dealt with? What are the *physical properties* that affect the measure? At what *level of detail* should the system be measured? How *intuitive* and *appealing* is the measure? These questions relate to three major properties characterizing the measure: *(i) conformity, (ii) granularity and (iii) intuitive appeal and rationality.* In this section we discuss these properties, to be used later in examining the fairness measures proposed recently in the literature.

2.1 Conformity to the general concept of social justice

For many people, fairness perception is very intuitive, almost instinctive. Why do people consider FIFO to be most fair in many situations? The answer is that it is "naturally" fair. Thus, approaches towards fairness of the queue are mostly based on pragmatic principles, e.g. seniority must be respected, or, customers requiring little should get priority (Short versus Long), or, waiting time should be in proportion to the service required. These pragmatic approaches are not necessarily directly based on an abstract general conception offered by "deep thinkers". Nevertheless, also the general conception worked out by the deep thinkers (mostly philosophers) emerges from the same "natural" pragmatic cultural attitudes of the society, and is a product of much discussion and lengthy discourse by these thinkers who, frequently, represent some of the finest minds of the society. The underlying principle of a fairness measure should conform to the general cultural perception of social justice prevailing in the particular society, either directly or indirectly. If it doesn't, its acceptance and usefulness may be deterred by inconsistencies and "surprises" in the form of counter-intuitive and unaccepted results.

2.2 Granularity

At what granularity level should the fairness performance metric conform to the underlying fairness principle? Our conclusion is that conformity is desirable on all three granularity levels of the system: *(i) the individual customer level, (ii) the scenario level (scenario is defined as a* sample *path of the stochastic process), and (iii)* the *system level.* The measure should be useful in assigning a consistent and meaningful fairness (or unfairness) value to each individual customer, to each possible scenario and to the system as a whole. The following is a more precise description of the three levels:

(i) ***Individual Customer (Job) Unfairness (Discrimination):*** This is a quantity attributed to the individual job (customer). It represents the deviation of the treatment given to the

customer, in a particular scenario, from the absolutely fair treatment as defined by the underlying fairness principle of the measure.

(ii) ***Scenario (Sample path) fairness:*** A summary-statistic that summarizes the discrimination as experienced by a (finite or infinite) set of jobs in a particular scenario (a sample path).

(iii) ***System fairness***: A summary statistic of a probabilistic measure (e.g. expected value or variance) of the performance as experienced by an arbitrary job, when the system is in steady state. This can be extended to a similar measure for transient behavior of the system

Addressing fairness at all three levels is similar to the addressing of the waiting time measure, which can also be evaluated at these three levels. It should be noted that queueing theory has dealt explicitly mainly with the third type of quantity (expected delay or its variance), as the other quantities are somewhat trivial in the context of customer delay. In the context of fairness, it is nonetheless essential to make explicit use of the individual and scenario quantities as well, since humans can feel them better and associate with them better than with the third quantity. This is important to building confidence in the fairness measure, which is somewhat abstract, non-tangible and difficult to feel.

In deriving the system fairness one may take two different approaches for dealing with the stochastic nature of the system:

1. ***Fairness of actual measures:*** This approach first computes the fairness for all situations in the system, and then uses some summary statistics function (e.g. the *max* operation or some type of *expectation)* to yield the system fairness measure. Thus, the approach compares the *actual performance measure* (as opposed to comparing the expected performance measure) observed by the individuals.
2. ***Fairness of the mean****:* This approach classifies the customers into classes and computes the expected performance (e.g. expected delay) of each of the classes; then all these expected values are compared to each other (by some summary statistics function, e.g. the *max* operation) to yield a measure of system fairness. Thus, in the comparison stage the entities that are compared to each other are the expected performance measures of the individuals and not the actual performance of the individuals.

To illustrate the difference between these approaches and the importance of granularity, consider any "pie division" problem, for example, a bonus b divided by an employer among n equally deserving employees. One approach is to consider the *actual* bonuses, $\{b_1, b_2, \ldots, b_n: b_1+b_2+\ldots+b_n=b\}$, given to the employees, compare them to each other and then use a summary statistic to summarize them. Since all employees are equally deserving the *absolutely fair* slicing of the pie is into equal shares, namely, $b_1=b_2=\ldots=b_n=b/n$, then the discrimination (positive or negative) of employee i is expressible as (b_i-b/n), namely, the deviation from absolute fairness. We note that the sum of discriminations is always zero, since this is a zero-sum situation; if one employee gets more it is taken away from other employees. The unfairness of the scenario can then be given by the averaged absolute values of the individual discriminations, or more convenient for analysis, by the averaged squared discrimination, $\Sigma(b_i-b/n)^2/n$.

Suppose now that the employer decides to use a probabilistic mechanism for slicing the pie. As a result, the bonuses are random variables $B_1,B_2,\ldots,B_n$ summing to b. The unfairness of the system, resulting from the unfairness of the scenario (synonym to realization) when using the

first approach, is given by $\Sigma E[(B_i-b/n)^2]/n$. In this approach the unfairness is defined and computable for all three granularity levels. An alternate approach is to use the concept of *fairness of the mean*. In this, alternate approach the distribution of the pie is fair if $E(B_i)= b/n$ for $i=1,2, ...,n$. This yields a *criterion*[6] for classifying systems into fair ones and unfair ones. What it really classifies is the probabilistic mechanism, or the lottery. The criterion is not applicable at the individual level or the scenario level. If one tries to apply it to a scenario, it will, in most cases, classify all possible scenarios (realizations) of a "fair" lottery as being unfair. To further illustrate, suppose the employer takes an "all or none" approach, by which it will be decided by lottery to grant one of the employees all the bonus money b, and all others get nothing. The *fairness-of-the-mean* principle will classify this system as fair, provided that the lottery gives even odds to all employees. Nevertheless, all possible scenarios (realizations) will be classified as unfair by the same criterion. In a gambling environment, where participants are psychologically prepared and willing to gamble, this lottery will be considered to be fair. It is very doubtful that employees will view it as gamblers do. Employees (and likewise customers in a queue) are individuals and experience unfairness individually and personally. Most of them are likely to view the implementation of this lottery as highly unfair, when compared to a deterministic even division of the bonus money. Putting it differently, the employees, unlike the fabled statistician, are wary of the possibility of drowning in the lake, notwithstanding its six inches *mean* depth. The first approach recognizes the difference in unfairness between the fair deterministic division of the bonus money and the "fair" division by use of the lottery. It assigns system unfairness value of zero to the former and $b^2(n-1)/n^2$ to the later, ($(n-1)/n^2$ if normalized by setting b=1).

2.3 Intuitive appeal and rationality

Producing intuitively acceptable results is a highly important, maybe the most important, property expected of a fairness measure. Surprising results, whose disagreement with intuition cannot be rationally and convincingly explained, are most likely to be rejected. A measure producing such "surprises" is not likely to achieve wide acceptance and might be viewed, instead, as an interesting curiosity. A good measure is not supposed to invent "new" fairness; it is supposed to quantify the prevailing widely accepted conceptions of fairness. Reactions like "What's so interesting about this measure? I knew intuitively that discipline ϕ is the fairest, without the help of the measure", are perhaps the strongest proof of the validity of the measure. The question often is how *much* more fair is discipline ϕ as compared to other disciplines, under various operating conditions.

In this section we propose four, intuitively based, simple tests for the validity of a measure. These do not suffice to label a measure as valid, rather, not passing them is a red light that the measure is questionable.

Two fundamental quantities determine the queueing process and the job scheduling decisions. These are the arrival epochs and service times, a_i, s_i of the customer C_i, $i=1,2,\ldots$. As our goal is to focus on the fairness of the queue and neutralize other external parameters, we will deal with

[6] The approach can also yield a measure of unfairness in the mean. Suppose the lottery is such that $E(Bi) \neq b/n$ for some values of i. Then $\sum[E(B_i)-b/n]^2/n$ can be used as a measure of unfairness in the mean.

these variables only. (And, as discussed earlier, we will not account for external parameters, such as neediness of customers, payments made by customers to obtain preferential service, or a gold/silver/bronze classification of customers.) Since these quantities are the only remaining ones determining the queueing and scheduling process, they also serve as the fundamental variables for determining scheduling fairness. For convenience of presentation, we use the terms *seniority*, and *service requirement*. The seniority of J_i at epoch *t* is given by $t - a_i$. The service requirement of J_i is s_i. One may recall that *seniority* and *service-requirement* were in the heart of the dilemma in the Short vs. Long scenario.

It is natural to expect that a "fair" scheduling discipline will give preferential service to highly senior jobs, and to low service-requirement jobs. This can be stated formally in the following two tests:

1. *(Weak) Service-requirement Preference Test*: If all jobs in the system have the same arrival time, then 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.
2. *(Weak) Seniority Preference Test*: If all jobs in the system have the same service times, then for jobs J_i and J_j, residing concurrently in the system, if $a_i < a_j$ then it will be more fair to complete service of J_i ahead of J_j than vice versa.

A stronger form of the preference tests is as follows:

3. *Strong Service-requirement Preference Test:* 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.
4. *Strong Seniority Preference Test:* 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.

The seniority preference test is rooted in the common belief that jobs arriving at the system earlier "deserve" to leave it[7] earlier. The service-requirement preference test is rooted in the belief that it is "less fair" to have short jobs wait for long ones. It should be noted that when $a_i < a_j$ and $s_i > s_j$ (the Short vs. Long case) the two principles conflict with each other, and thus the relative fairness of the possible scheduling of J_i and J_j is likely to depend on the relative values of the parameters.

One may view these two preference tests as two axioms expressing one's basic belief in queue fairness. As such, one may expect that a fairness measure will satisfy these tests. A fairness measure is said to satisfy a preference test if it associates higher fairness values with schedules that are more fair. A formal definition is given next:

[7] An alternative view to "leave it earlier" is "enter service earlier". These two alternative concepts are equivalent when service times are identical and service is uninterruptible. The latter concept might lead to difficulties in the case of service interrupting scheduling (e.g. any preemptive regime).

Definition: Consider jobs J_i and J_j, requiring equal service times and obeying $a_i < a_j$. Let π be a scheduling policy where the service of J_i is completed before that of J_j and π' be identical to π, except for exchanging the service schedule of J_i and J_j. A fairness measure is said to satisfy the strong seniority preference test if the fairness value it associates with π is higher than that it associates with π'.

Similar definitions can be given to the service-time preference test and to the weak-versions of the preference tests.

It is easy to see that if a fairness measure satisfies the *strong preference test* (either Service-requirement or Seniority) then it must also satisfy the corresponding *weak preference test*.

To illustrate the preference tests in the context of scheduling policies we review several common policies and examine whether they follow the preference tests. A formal definition is:

Definition: A scheduling policy π is said to satisfy the strong seniority preference test if for every two jobs J_i and J_j, requiring equal service times and obeying $a_i < a_j$, π completes the service of J_i ahead of that of J_j.

A similar definition can be given for the strong service-time preference test and for the two weak preference tests.

Using these definitions, one can classify common scheduling policies as follows:

a. ***FIFO***: The *First-In-First-Out* scheduling satisfies the strong seniority preference test. On the other hand, since it gives no special consideration to shorter jobs, it does not satisfy the service-time preference tests (weak or strong).

b. ***LIFO*** and ***ROS****:* The *Last-In-First-Out* and *Random Order of Service* policies do not satisfy the seniority preference test (either strong or weak). Furthermore, neither do they satisfy the service-time preference test.

c. ***SJF*** and ***LJF:*** The *Shortest Job First (SJF)* satisfies the strong service-time preference test. Nonetheless – it does not satisfy the seniority preference test (both strong and weak). The *longest Job First (LJF)* satisfies none of the tests.

d. ***PS*****:** The *Processor Sharing* policy satisfies both the strong seniority preference and the strong service-time preference tests.

e. ***FQ:*** *Fair Queueing*, which is the non-weighted version of Weighted Fair Queueing (Parekh (1992) and Parekh and Gallager (1993)), serves the jobs in the order they complete service under Processor Sharing (unless some of the jobs are not present at the server at the time that the service decisions must be taken). This property and the fact that PS satisfies both of the strong preference tests, imply that FQ satisfies both the strong seniority preference and the strong service-time preference tests.

3 A review of proposed fairness measures and their properties

Analytic treatment and quantification of queue fairness have been quite limited in the literature, and been addressed only very recently. Five references that propose measures, a criterion or an approach for quantitative evaluation of fairness of queues are Avi-Itzhak and Levy (2004), Raz,

Levy and Avi-Itzhak (2004), and Sandman (2005), who propose measures; Wierman and Harchol-Balter (2003), who propose a criterion, and Gordon (1987), who proposed an approach to evaluating fairness of the queue. The modeling dilemma of seniority versus service requirement seems to be at the heart of these queue fairness-modeling attempts: The approaches proposed in Gordon (1987) and in Avi-Itzhak and Levy (2004) center on the *seniority* factor. In contrast, the approach proposed by Wierman and Harchol-Balter (2003), focuses on the *service-requirement* factor. Sandman (2005) proposes to consider both *seniority* and *service requirement*, and lastly, Raz, Levy and Avi-Itzhak (2004) focus on neither of them and choose to focus on fair *resource allocation*[8], directly conforming to the general conception of social justice. In this section we review these publications and examine their properties in light of the discussion of expected properties given in Section 2.

3.1 Seniority Based Fairness: Order Fairness

3.1.1 Skips and Slips: An Approach for Fairness Evaluation

An approach for evaluating fairness, based on seniority through counting of "skips" and "slips" was proposed by Gordon (1987) in a doctoral dissertation. The approach aims at quantifying the violation of social justice due to overtaking in the queue. The underlying rationale is that FIFO is just and customer overtaking causes injustice.

The approach defines two types of overtaking events experienced by a tagged customer in a queueing system: 1) A *Skip* – when the tagged customer overtakes another customer (namely, it completes service before a customer that arrived ahead of it), and 2) A *Slip* – when the tagged customer is overtaken by another customer. Gordon (1987) suggests that counting the number of skips and slips can provide an indication to the amount of injustice and focuses on analyzing these counts; nonetheless – it does not deal with how to use these as the basis for a fairness measure.

Several systems are studied under that approach: 1) Two M/M/1 systems in parallel, 2) Two M/M/1 systems in parallel, where the tagged customer (and only him) uses the "join the shortest queue" strategy, 3) The multi server system, M/M/m, and 4) The infinite-server system, M/G/∞. For these systems the probability laws of the number of skips and the number of slips experienced by an arbitrary customer (denoted N_{SKIPS} and N_{SLIPS}, respectively) are derived. Interesting results are: 1) For every system $E\{N_{SKIPS}) = E(N_{SLIPS})$, 2) For most systems the distributions of the two variables differ from each other, 3) Only one system is found by the author where the distributions equals each other, the M/G/∞ system where the service time distribution is symmetric around its mean, 4) Using the "join the shortest queue" strategy by the tagged customer, when no one else uses it, reduces the number of slips and increases the number of skips he/she experiences.

[8] We note that the three physical quantities forming the basis for these measures, namely arrival time (*seniority)*, *service requirement* and *resources,* are the fundamental quantities on which every queueing system is based and are used in the queueing-theory notation of X/Y/z (e.g. M/G/1) to characterize a queueing system.

In the following this approach is viewed in the light of the three basic properties discussed in Section 2:

1. **Conformity:** The underlying principle is pragmatic, namely, seniority merits priority. Nonetheless, the approach is not fully sensitive to seniority, since it assigns equal weight to skips (and slips), regardless of the relative seniority of the involved customers: If customer C_j skips customer C_k the same weight will be accounted for regardless of whether C_j arrives a second behind C_k or an hour behind C_k. Also, it does not consider service requirements at all, and thus it may apply mainly in systems where seniority is the most important factor, e.g. identical deterministic service times (and possibly exhaustible-servers systems with equally needy customers). In the case of non-equal service times the conformity of this principle to the conception of social justice at large might be questioned.
2. **Granularity:** Accounting for the number of skips and slips can be done at all three granularity levels, individual, scenario and system. How fairness at the system level can be measured, remains open in the work of Gordon (1987). In light of the fact that $E\{N_{SKIPS} - N_{SLIPS}) = 0$, a possible approach that comes to mind is to take $Var\{N_{SKIPS} - N_{SLIPS})$ as a system fairness metrics. How such a measure behaves and how it relates to the measure developed in Section 3.1.2 below is an open question.
3. **Intuitive appeal and rationality:** The approach is strongly intuitively appealing, as long as only seniority matters, since it is based on the concept that FIFO is most fair. Since no system measure was proposed or studied, the question whether it satisfies the basic tests is not meaningful.

3.1.2 A Seniority Based Fairness Measure

An order fairness measure, based on seniority, was studied in Avi-Itzhak and Levy (2004). The basic underlying model used in that study assumes that all service times are identical. In that context the major factor of interest is that of job-seniority. The study deals with a specific sample path of the system, and examines a realization π of the service order (that is, a feasible sequence of job indices reflecting the order of service), and with a fairness measure $F(\pi)$ defined on the service order. The paper assumes several elementary axioms on the properties of $F(\pi)$. The major axiom is:

***Monotonicity of F() under neighbor jobs interchange*:** If two neighboring jobs are interchanged to modify π and yield a new service order π' then $F()$ increases if the interchange yields advancing the more senior of the two jobs ahead of the less senior job, and it decreases if the interchange advances the less senior job ahead of the more senior job. If the seniority of the interchanged jobs is the same – $F()$ is not affected by the interchange.

The additional axioms deal with 2) Reversibility of the interchange, 3) Independence on position and time**,** and 4) Fairness change is unaffected by jobs not interchanged**.**

The reader may recognize that the core axiom of this approach, Axiom 1, is simply a mathematical form to express the *seniority preference test* presented in the previous section.

The results derived in Avi-Itzhak and Levy (2004) show that for a specific sample path the quantity $c\sum_i a_i\Delta_i + \alpha$, where Δ_i is the *order displacement* of customer C_i (number of positions C_i is pushed ahead or backwards on the schedule, compared to FIFO), and where $c>0$ and α are arbitrary constants, satisfies the basic axioms. This quantity is the unique form satisfying the axioms applied to any feasible interchange (not necessarily of neighbors).Under steady state this quantity is equivalent to the *variance* of the *waiting time* (with a negative sign). Thus, when all *service times* are *identical* the *waiting time variance* can serve as a surrogate for the *system's unfairness measure*.

In the following this measure is viewed in the light of the three basic properties discussed in Section 2:

1. **Conformity:** The underlying principle is pragmatic, namely, seniority merits priority. In the case of equal service times, for which it is proposed, it can be considered to conform to the basic conception of social justice, providing that the disutility of the wait is part of the pie. One possible way of extending this concept to the non-constant service times situation, is to require that the waiting disutility be divided in proportion to the slice of the resource received by each customer. (See discussion in Section 1.2.2).
2. **Granularity:** The measure is defined, and is applicable, at all three granularity levels; the individual customer level, the scenario level and the system level.
3. **Intuitive appeal and rationality:** The measure is strongly intuitively appealing as it is based on the concept that FIFO is most fair, and since service times of all customers are the same, the Short-versus-Long conflict does not exist.
   1. The measure satisfies the strong *Seniority Preference test* (Section 2.3). This can be verified by recalling that the unfairness function for a sample path is given by $\sum_i a_i\Delta_i$ and by examining the change of this function due to the interchange of J_i and J_j.
   2. When all service times are identical, the fairest policy in the class of work conserving and uninterrupted service policies is FIFO. The most unfair policy under these conditions is LIFO.
   3. The measure, if used for non-equal service times, does not satisfy[9] the *Service-requirement Preference test* (Section 2.3). If one uses the variance of waiting time as a measure of unfairness, then there are cases where it is more fair to serve a long job ahead of a short job. For example consider a system with two jobs only, J_1 and J_2 whose service times are $s_1 = 1, s_2 = \varepsilon \to 0$. Serving the longer job J_1 first leads to a waiting time variance that approaches 0 while serving the shorter job J_2 first leads to a waiting time variance that approximately equals 1/4.

3.2 An Expected-Slowdown Based Fairness Approach

Slowdown, S(x), is defined in computer related queueing publications as the conditional response time divided by the conditioning service length: $S(x) \stackrel{def}{=} T(x)/x$, where $T(x)$ is the

[9] In fact, it might not be appropriate to examine this test as the measure is built for equal service-time situations.

response time experienced by a customer whose required service time is of size x. Wierman and Harcol-Balter (2003) propose a criterion for classifying M/G/1 disciplines into three classes (based on earlier work on slow-down presented in Bender, Chakrabarti and Muthukrishnan (1998), Bansal and Harchol-Balter (2001) and Harcol-Balter, Sigman, and Wierman (2002)):

- A scheduling policy is said to be fair for given load and service distribution if $E[S(x)] \le 1/(1-\rho)$ for all values of x, where $\rho<1$ is the system's load ($\rho=\lambda E(s)$, where λ is the arrival rate and s is the service time).
- A service policy is *always fair* if it is fair under all loads and all service distributions. A service policy is *always unfair* if it is not fair under all loads and all service distributions.
- Other policies are *sometimes unfair,* meaning fair under some loads and distributions and unfair under others.

Although, this is a criterion, in contrast to a measure which assigns a numerical value to each M/G/1 discipline, we include it here since it raised interest in the computer science queueing related community and it resembles the pragmatic principle of waiting and service proportionality, mentioned in Section 1.2.

The work of Wierman and Harcol-Balter (2003) analyses a large set of scheduling disciplines under the M/G/1 setup, and classifies them into the above three classes.

In the following this approach is viewed in the light of the three basic properties discussed in Section 2:

1. **Conformity:** The criterion is based on *fairness-of-the-mean* approach. It is not clear how this criterion conforms to the general perception of social justice. The paper proposing it does not provide an explanation of the underlying rationale. Rather it sends the reader to earlier publications. The impression is that a discipline is axiomatically assumed to be fair if $E[S(x)] = 1/(1-\rho)$. To quote Harcol-Balter, Sigman and Wierman (2002): *For any given load $\rho<1$, under PS scheduling, all jobs have the same expected slowdown; hence PS is "fair".* Our impression is that the criterion is intended to classify a discipline P as fair if it is at least as efficient as PS for all values of service requirement, x. Namely, $E(T_P(x)) \le E(T_{PS}(x))$ for all values of x, where $T_P(x)$ and $T_{PS}(x)$ are the conditional response times under P and under PS, respectively. P is always fair if the inequality holds for all loads ρ and all distributions of s. It seems that it is not as much the proportionality (in the mean) that counts, as is the efficiency. Theoretically P may be always fair without satisfying the proportionality. In justifying the use of $1/(1-\rho)$ the authors show that there exists no policy P such that $E[S_P(x)] = c < 1/(1-\rho)$, where c is some constant. That is, no P that satisfies $E[S_P(x)]=$ *constant* is more efficient than PS, in contrast to being at least as efficient as PS.

   An additional rationale for the use of the expected slow-down as a fairness criterion was offered to us in personal communications. In systems where the customer does not see other customers (such as in many computer systems) the customer can view his response time only relatively to his service requirement, and not relatively to other customers concurrently served with him in the system. Note, however, that to adopt this rationale the customer must, somehow, be able to relate to his *expected* response time and must not care about how the blind queue internally schedules jobs (see the discussion on blind queues in Section 1.3.1);

This philosophy can be subject to difficulties when the internal information about the blind queue becomes, sooner or later, available to customers.

2. **Granularity:** The criterion is based on expected values. As such it yields to *system fairness* analysis for a wide class of M/G/1 disciplines. However, customers are individuals and they experience unfairness individually (see the bonus example in Section 2.2). To this end - the criterion is not applicable to classifying unfairness to individuals or unfairness of a scenario. It applies only to the steady-state M/G/1 process, and does not allow deciding the relative fairness rankings of individual disciplines, in case they belong to the same class (e.g. always fair). To clarify this issue, we will explore it a little more extensively, by examining the 'always fair' policies.

Wierman and Harcol-Balter (2003) classified the LIFO-PR and PS policies as *always unfair*. They posed as an open question for research whether there exist other *always fair policies*. Below we describe an indefinitely large class of policies that are *always fair*. It is also shown that the differences between them with respect to slow-down can be very drastic. Consider the class of M/G/1 time-sharing queues with finite number of service positions, (Avi-Itzhak and Halfin (1987)) described below:

> The queue is ordered and there are *r* service positions and an unlimited number of waiting positions. When there are *n* jobs in the system they are in positions 1, 2,. . ., *n* and a proportion $\varphi(i,n)$ *)* of the service rate is directed at the job in position *i, (i* = 1, 2,. . ., *min(r, n))* where $\sum_{i=1}^{\min(r,n)} \varphi(i,n) = 1$, *n* = 1, 2, ... A newly arrived job which encounters *n* in the system, *n* = 0, 1, 2,. . ., is assigned to service position *i* with probability $\varphi(i,n+1)$. The other *n* jobs present in the system at that time are re-ordered in accordance to some arbitrary service-independent rule. (A re-ordering rule is service-independent if it involves permutations based only on the knowledge of the number of jobs, *n,* and their positions in the queue). When the processing of a job is completed it departs, and the remaining jobs are instantaneously re-ordered in accordance to some arbitrary service-independent rule. Preemptions due to newly arriving jobs or due to reordering do not result in loss and when a preempted job re-enters service its processing resumes from the point of the most recent interruption.

It can be verified, see Yashkov (1980), that all members of this class possess the equilibrium properties of symmetric queues as defined by Kelly (1979). Thus, for all members of this class the expected delay and response times of a job requiring *x* units of service time are given respectively by $E(W(x)) = x\rho/(1-\rho))$ and $E(T(x)) = E(W(x)) + x = x/(1-\rho)$.

By the proposed criterion all members of this class are *always fair*. We note that if $\varphi(i,n) = 1/i, i = 1, 2,..., \min(r,n)$, we get the PS queue when $r \to \infty$ and an indefinitely large class of LCFS-preemptive queues when *r =1*. In this class the LIFO-PR[10] (called the stack by Kelly (1979)) is the one where a displaced customer is always placed in position 2 and all customers in the waiting positions are moved one position back, and the one in position 2, if

[10] The reader should be careful and note that we distinguish here between LCFS-preemptive (to denote the family) and LIFO-PR (to denote the specific policy); In the literature this specific policy is called sometimes LIFO-PR and sometimes LCFS-PR.

not empty, is always the one to enter service when a customer departs. Let the conditional delay be denoted by $W^f(x)$ for LIFO-PR, by $W^l(x)$ for LCFS-PR where the displaced customer is placed at the end of the line, and by $W_\infty(x)$ for the PS case. Avi-Itzhak and Halfin (1987) show that $Var(W^f(x)) \geq Var(W^l(x))$ and $Var(W^f(x)) \geq Var(W_\infty(x))$. They conjecture that the variance of the conditional delay is decreasing in *r* and therefore $Var(W(x)) \geq Var(W_\infty(x))$, where $W(x)$ is the conditional waiting time in any arbitrary policy in this class. For the case of exponentially distributed service times they show that

$$\frac{Var(W^f(x)) - Var(W^l(x))}{Var(W^f(x))} = \frac{\rho[1-\exp(-\lambda(1-\rho)x)]}{\lambda(1-\rho)x} \geq 0.$$

For a fixed value of *x*, this expression increases with ρ and goes to ∞ when ρ→1. For the exponential case they also show that $Var(W^l(x)) \geq Var(W_\infty(x))$. The variance reduction obtained when preempted jobs are placed at the end of the line, or when PS is used, is very significant when the traffic intensity is high. For example, if λ=0.9, ρ=0.9 and *x*=0.5 we get $Var(W^f(0.5)) = 90.4Var(W_\infty(0.5))$ and $Var(W^f(0.5)) = 18.5Var(W^l(0.5))$. If the deviation from fairness is defined as $W(x) - x\rho/(1-\rho)$ and the variability of this deviation reflects the unfairness, then for customers requiring service time of 0.5 the LIFO-PR is 90.4 times <u>*more unfair*</u> than the PS schedule, and the relation above implies that this ratio can become infinitely large. Yet the criterion classifies both as *always fair*.

We are not aware of any policy outside the family of symmetric queues, which is *always fair*. This might be an interesting question for future research.

3. **<u>Intuitive Appeal:</u>** The criterion classifies as *always unfair* all conservative policies that are: (*i*) Non-preemptive non-size based (e.g. FIFO, LIFO and ROS); (ii) Preemptive size-based (e.g. preemptive-shortest-service-first), or (iii) Age based (age of a job is defined as service already received). It classifies as *sometimes unfair* all conservative policies that are: (i) Non-preemptive size-based (e.g. shortest-service first); (ii) Preemptive shortest remaining processing (all other remaining-processing based are either *always unfair* or *sometimes unfair*). As stated earlier, it classifies as *always fair* all symmetric queues (including LIFO-PR and PS). The tests of fairness defined in Section 2.3 do not apply, since the criterion does not have the necessary granularity. The criterion is not likely to intuitively appeal to humans, who will have great difficulty to accept, in the context of daily life, that LIFO-PR is *always fair* while FIFO is *always unfair*, even for M/D/1. An investigation of the measure suggested in the granularity discussion above is worthwhile.

3.3 A Service-requirement and Seniority Combination Based Fairness

A fairness approach based on accounting both for service-requirement and seniority (as a combination) is offered in Sandmann (2005). Similar to Gordon (1987) (see Section 3.1.1) the approach aims at counting events of fairness violation. In addition to the seniority violation events offered in Gordon (1987), the author proposes to count also events of "size" violation. More specifically, a tagged customer *C*, in this approach, can be subject to two types of "discriminating" events: 1) An *overtaking event*, in which *C* is overtaken by another customer.

This is exactly identical to the slip-event (see Section 3.1.1). 2) A *Large Job event* - this event occurs *if* upon the arrival of *C* to the system it finds there *C'* whose residual service is greater than or equal to the service requirement of *C*, and (later) *C'* departs from the system ahead of or concurrently with *C*. Let N_i^{over} and N_i^{large} be the number of these events, respectively, experiences by customer C_i, *i=1, 2, ...* The discrimination frequency[11] of C_i is defined to be $DF_i = N_i^{over} + N_i^{large}$, and the discrimination frequency of a sample path π is defined as $DF(\pi) = \sum_i DF_i$. Let N^{over} and N^{large} be the number of discrimination events experienced by an arbitrary customer when the system is in steady state. The system unfairness under steady state can be defined as $E(N^{over} + N^{large})$.

Sandmann (2005) showed that it satisfies both strong preference tests. The question of how to derive the expected discrimination of an arbitrary customer is not addressed. To yield this measure note that $E(N^{over})$ can be taken from the analysis in Section 3.1.1); the analysis of $E(N^{large})$ remains as an open issue for research.

In the following this measure is viewed in the light of the three basic properties discussed in Section 2:

1. **Conformity:** The underlying principle is pragmatic, namely, seniority and smaller residual service time merit priority. Nonetheless, the accounting for the "overtaking" and "large" events raises difficulties in three aspects: 1) Overtaking events are not fully sensitive to seniority differences (see the remark regarding skip events in Section 3.1.1); 2) Similarly, large events are not sensitive to the difference in remaining service requirements; 3) How to weigh "overtaking" events versus "large" events is not clear; the equal weight used by the author (using the simple sum of these variables) might be arbitrary.
2. **Granularity:** The measure applies at all levels of granularity. These are given by DF_i for the individual customer, $DF(\pi)$ for a sample path and the expected value of the *DF* measure taken for an arbitrary customer in equilibrium.
3. **Intuitive appeal and Rationality:**
   1. The measure satisfies *both* the strong Seniority Preference test and the strong Service-Requirement Preference test. . It is worth noting that this is the only measure, of the measures reviewed here, which satisfies both strong tests.
   2. While satisfying the two strong tests, the measure seems to behave non-rationally in other cases; this is caused by the difficulty of combining the overtaking and large events into one measure. To demonstrate such a difficulty consider the following case: Customer C_i whose service time is 100 seconds arrives at the system. Customer C_j whose service time is 99.9 seconds arrives an hour later. Now if C_i is served first, C_j experiences a "large" event (and C_i experiences no event). If, on the other hand, C_j is served first, C_i experiences an "overtaking" event (while C_j experiences no event). That is, the

[11] The term "discrimination frequency" is used in Sandmann (2005). The term "discrimination count" might be more appropriate in this context.

system (and sample path) measures of fairness are insensitive to the order of service, and the large extra seniority of C_i is washed by the infinitesimal service-requirement advantage of C_j. While it may be possible to assign weights to the overtaking and large events, based on seniority differences and job-size differences, the remaining challenge is how to determine the trade-off between the total weights of the two event types. This issue calls for further research.

3. Also, it seems that the weak residual-service inequality used to define the *Large* events causes the measure to be unintuitive in the case of G/D/1 with FIFO discipline. This system is expected to be very fair. Nonetheless, using the weak inequality each arriving customer experiences as many as *N-1 Large* events upon its arrival, where *N* is the number of customers it sees upon its arrival. We believe that this can be fixed by making *strict* the residual-service inequality in the definition of *Large.* Nonetheless, we do not know whether such a modification will adversely affect item 1 above.

3.4 A Resource Allocation Based Fairness

A Resource Allocation Queueing Fairness Measure (RAQFM) was introduced in Raz, Levy and Avi-Itzhak (2004). The measure is directly based on the general conception of justice requiring equal distribution of resources among all equally needy members. The measure accounts, indirectly, for both seniority and service-requirements, thus offering a solution to the Short versus Long conflict, based on a basic principle of social justice. The method applies to multiple servers (Raz, Levy and Avi-Itzhak (2005)), but for the sake of presentation we will focus mostly on the single server system.

The basic underlying principle is that at any moment in time all customers present are entitled to equal shares of the resource, namely, at every epoch *t* at which there are *N(t)* jobs (customers) present in the system, each is entitled to *1/N(t)* of the server's serving rate. This is called the temporal *warranted service rate* to be given to each customer at that epoch. The overall warranted service of C_i (customer *i*) is given by integrating this value over the duration that C_i stays in the system. Subtracting this warranted service from the granted service (which is the service granted to C_i, namely its service time, s_i) yields the *discrimination* of C_i, denoted $\delta_i = s_i - \int_{a_i}^{d_i} \frac{1}{N(t)} dt$, where d_i is the departure epoch of C_i. Note that the discrimination may be positive or negative. Taking summary statistics over all discriminations experienced by the customers yields an unfairness measure for the system. This measure can apply to a specific scenario (sample path), to yield the unfairness of that path. Similarly, taking expectations of this measure over all sample paths yields the system unfairness. One of the basic properties of the discrimination function is that it is a zero-sum function (namely the total discrimination in the system, at every epoch, is 0). Thus, the expected value of discrimination is meaningless, and the

proper summary statistics is the second moment[12] (or variance) or expected absolute value of discrimination.

The measure yields to exact analysis (via numerical procedures) for the family of multiple-server Markovian (including M/Phase-type/m type and in particular M/Coxian/m type) queues. It is an open subject for research whether (and how) it yields to mathematical analysis for M/G/1 type systems and to systems with arbitrary service times, at large. It does yield, for example, to exact mathematical analysis of the M/G/1 LIFO-preemptive queue (Brosh, Levy and Avi-Itzhak (2005)).

1. **Conformity:** The underlying principle of the measure conforms directly to the basic conception of social justice.
2. **Granularity:** The measure is defined, and is applicable, at all three granularity levels; the individual customer level, the scenario level and the system level.
3. **Intuitive appeal and rationality:** This measure is not based on a pragmatic intuitive principle, but rather on a general conception of social justice. As such it may not be intuitively appealing at a first glance. Therefore, an extensive examination of its properties under various systems and conditions, and their agreement with intuitive appeal, is called for. Below we review part of these properties, which are remarkably in line with intuition and rationality.

   I. Basic Properties – single server (Raz, Avi-Itzhak and Levy (2004a)):

   1. The measure satisfies the *Strong Seniority Preference test* for work conserving and non-preemptive service policies.
   2. When all service times are identical, the fairest policy in the class of work-conserving and non-preemptive policies is FIFO. The most unfair policy under these conditions is LIFO. This holds also for general-independent service times when the service order is independent of the service times.
   3. The measure satisfies the weak *Service-requirement Preference test* for work conserving and non-preemptive policies. Nonetheless, it does not satisfy the strong version of this test, as there exist some counter examples.

   II. Effect of Service Time Variability on Fairness:

   Since fairness of the queue is mainly obtained by a desirable preference balance between a customer's seniority and his service requirement, service time variance is intuitively a major factor. The larger the variance, the more prevalent are situations with an acute conflict between seniority and service length. At the other extreme, where all service times are the same there is no conflict at all. As the variance of service times increases we expect the unfairness of all non-preemptive disciplines to increase. On the other hand, LIFO, which is extremely unfair, may intuitively be more fair than FIFO once we allow preemption, i.e. FIFO may be less fair than LIFO-PR, if the variance of service times is large. The intuitive reasoning behind this is that a customer with very long service time is likely to be preempted by one with shorter service time, thus achieving a better balance between seniority and service time. This is expected to be more pronounced when traffic is light and seniorities are

[12] In this case the units of the measure are time squared (such as delay variance). One can take the square root of it to make the units equivalent to those of mean delay.

relatively small. In heavy traffic seniorities are very large and therefore weigh more against shorter service requirements (see the Short vs. Long episode).

The values of the RAQFM measure agree surprisingly well with these intuitive expectations. This is demonstrated in Figure 3 and Figure 4 where the unfairness in systems with small service time variability ($\gamma_S = 1/10$), and with large service time variability ($\gamma_S = 10$) is depicted, respectively. These results are taken from Brosh, Levy and Avi-Itzhak (2004) who studied the effect of service time variability on the fairness of the queue, in the $M/G_{coxian}/1$ case, using RAQFM.

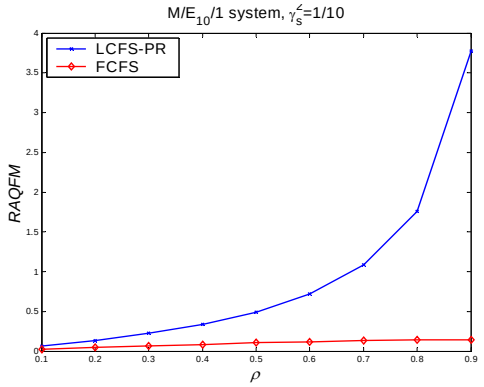

Figure 3: Unfairness (RAQFM) for low variability service time, (Coeff. Variation $\gamma_S = 1/10$)

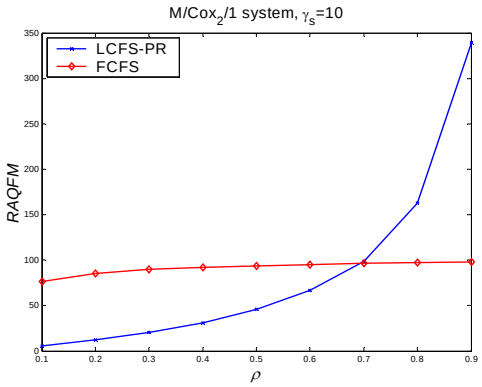

Figure 4: Unfairness (RAQFM) for high variability service time, (Coeff. Variation $\gamma_S = 10$)

III. Fairness in Multiple Queues

A question that was discussed in a number of papers is how the fairness of the fairness of the common (single) queue compares with that of the multi (individual) queues. This is studied in Raz, Levy and Avi-Itzhak (2004), and some of those results for two-server systems are shown in Figure 5 and Figure 6.

As shown in the figures, the FIFO common (single) queue is less unfair than the separate (multi) FIFO queues, in both constant and exponential service times cases. Permitting jockeying from the head of a queue or its tail, when the other server is idle, increases the fairness of the separate queues configuration. Nevertheless the common queue is still the fairest. In these particular examples arriving customers join the individual queues randomly. The results about the common queue being more fair than the multiple queue are re-enforced by the findings of Rafaeli et al (2005) and those mentioned by Larson (1987). The authors also show that joining strategies, like join-the shortest-queue and round-robin joining, improve the fairness of the separate queues configuration, but the common queue still comes out to be the least unfair.

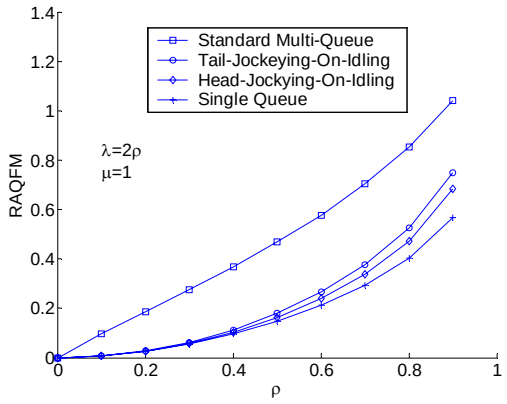

Figure 5: Unfairness (RAQFM) of four queue strategies under M/M/2 model

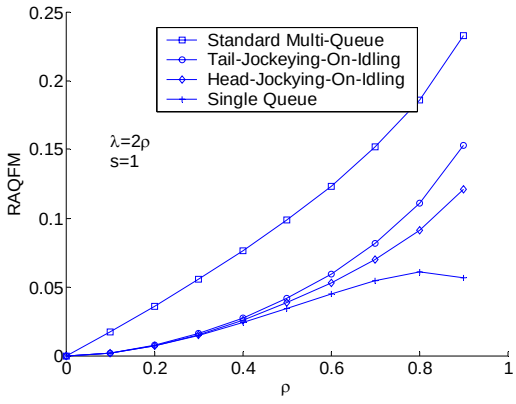

Figure 6: Unfairness (RAQFM) of four queue strategies under M/D/2 model

IV. Effect of Priority Classes on Fairness

The effects of priority classes on the RAQFM values of single server systems are investigated in Raz, Avi-Itzhak, and Levy (2004b). One interesting finding, agreeing with intuition, is that in a system with two classes, granting priority to the short-job class increases system fairness when the difference between the classes' service times is large, and decreases system fairness when the difference is small. In the latter case the effect of seniority violation dominates that of size preference, and in the former case the reverse is true.

4 Discussion: Measure Applicability and Future Research

Attempting to measure fairness in queues calls for fresh new approaches, deviating from the traditional ones of queueing theory, focusing - almost exclusively - on efficiency. This area is young, but gradually gaining recognition, and the importance of managing fairness *and* efficiency becomes more evident. As such it is a highly challenging, but also promising, area for researchers and practitioners alike. In this paper we exposed the existing, mostly very recent, research works on quantifying fairness of queues. In what follows we indicate several possible directions of future research. We recognize our inherent bias due to personal research involvement in this area and hope that fresh minds will generate new ideas and approaches, far beyond ours.

When trying to compare the measures and approaches suggested this far, the question of the degree of universal applicability comes to the surface, i.e. how wide is the range of systems each measure applies to, either in its original form or via a generalization. Wide applicability is one of the most important requirements of a measure, since if it is not applicable to many systems it may not be useful as a scale of reference. The order fairness measure presented in Section 3.2.1 applies only to the case where service times are equal or to the case where the servers are exhaustible and the major performance factor is getting the service at all. The slowdown criterion of Section 3.3 is defined for M/G/1 only. In both cases the degree of universality is quite limited. The skips-and-slips approach (Section 3.1.1), the Service-requirement and Seniority

Combination (SSCF) measure (Section 3.3) and the RAQFM (Section 3.4) are more universally applicable.

We demonstrate this using an example, raised by one of the referees, where service requirement and actual service time are not necessarily identical. Consider system (A) with two servers, a fast server with service rate μ_1 and a slow server with service rate μ_2, and let $\mu = \mu_1 + \mu_2$. How does the fairness of this system compare to system (B) with two identical servers of individual rates $\mu/2$? Intuitively (A) seems less fair than (B), since customers served by the faster server get preferential treatment.

For simplicity, assume that service requirement is identical to all customers, equaling 1 unit, and that the total service rate is $\mu = 2$. Further assume that the service rates are $m_1 = m_2/3 = 1/2$ in system (A) and $m_1 = m_2 = 1$ in (B). In this case the actual service times in (A) depend on where the customer is served (2 at server 1 and 2/3 at server 2), while in (B) the actual and required service times are identical, equaling 1. Suppose first that common waiting line with FIFO order is used. In system (B) no slips or skips take place, and thus both the skips-and-slips approach and the SSCF measure will find the system to be absolutely fair. In system (A) no slips or skips take place in the waiting line, but are possible during service, since a job admitted to the fast server may depart ahead of a more senior one admitted to the slow server. In fact, in the numerical example above, if traffic is heavy, every customer served by the slow server experiences two slips and two out of three customers admitted to the fast server experience one skip. One quarter of the customers are served by the slow server, which means that in heavy traffic the average number of skips plus slips per customer is approximately 1. Both the skips-and-slips approach and the SSCF measure will classify system (A) as more unfair than (B).

As for RAQFM, for multiple server systems (see Raz, Avi-Itzhak and Levy (2005)) it defines the discrimination of C_i as $\delta_i = s_i - \int_{a_i}^{d_i} \omega_i(t)\frac{1}{N(t)}dt$, where $\omega_i(t)$, is equal to the overall service rate available in the system at that epoch. This rate equals 2 (in heavy traffic), in both (A) and (B). In system (B) the length of stay in the system, $d_i - a_i$, is statistically the same for all customers. However, in system (A) a customer served by the slow server stays in the system, statistically, longer than a customer served by the fast server. One can then expect intuitively, the variance of discrimination, which is the unfairness measure, to be larger in system (A) than in (B). Intuition notwithstanding, this needs a mathematical proof.

Evidently, as indicated by the above example, the universality of proposed measures, and/or their generalization to wider settings and applications, is an important future research issue. What come to mind are systems such as queueing networks and queues with reneging and balking, among a richness of possibilities. In studying such generalizations, it is important that it will be mathematically tractable (to a reasonable extent) to afford exact analysis of the measure, or that it will at least be computationally feasible. It is also important to examine whether the results produced fit with basic intuition.

Another worthwhile direction is developing a measure of fairness based on the proportionality principle, namely, *waiting time of a job should be in proportion to the servers' time provided to it*. It is intuitively very appealing to require that customers who get more will also wait more.

This idea was addressed in Section 1.2.2 and discussed at some length in Section 3.2 where proportionality of the mean can be interpreted to be the principle underlying the slowdown based criterion. For M/G/1 queues we propose, in the granularity discussion of Section 3.2, a fairness measure based on individual discrimination of C_i, being defined as $W_i - \rho x_i /(1-\rho)$, where W_i and x_i are the waiting time and service requirement of C_i, respectively. This approach can also be extended to settings other than M/G/1 queues, where the proportionality constant is not necessarily $\rho/(1-\rho)$.

A third interesting problem is how to determine the trade-off between the total weights of the two event types of the SSCF proposed measure (Section 3.3). This issue involves directly resolving the fundamental dilemma of Short vs. Long. The RAQFM and the proportionality measure indicated above resolve this dilemma indirectly, by trying to adhere to an underlying basic fairness principle. No direct approach has been suggested so far.

The issues of how to account for different values of neediness or for economical factors, and of how to combine the issues of queue pricing with fairness remain open. Pricing/admission and scheduling received much attention in the queueing literature, starting perhaps with the paper of Naor (1969) who introduced the possibility of controlling the steady state length of a single-server queue by introducing prices for the service. A recent book (Hassin and Haviv (2002)) is dedicated to this subject. In this context a number of questions can be asked. One of them is how to account for the purchasing of queue privileges by a customer in the fairness framework. Such privileges can be purchased from the system (like the VIP example given at the end of Section 1.2.1) or by side-payments between the customers (consider again the Short vs. Long case and assume that Short offers Long a side payment to switch their places in the queue).

A simplistic approach to this question, which separates the issue of payments from that of fairness, may prevail in some cases. For example, in the case of the dedicated queue for VIP customers (who paid extra) one may claim that comparing VIP customers to non-VIP customers, by any of the fairness measures reviewed in this work, is irrelevant. The reason is that the VIP price is set ahead of time and each customer is free to choose whether to pay it and join the VIP class or not; thus, fairness should be evaluated only within each class. Similarly, one may claim that inter-customer side-payments can be separated from fairness as follows: The basic scheduling used by the system has some fairness value attributed to it (as described in this work). Once the schedule is set and customers negotiate queue positions with each other, in a "free market" mode, the *schedule change* is irrelevant to fairness, since it is done in free-choice mode. Of course, for this to hold, when Long and Short switch positions, other customers who are positioned between Long and Short must be involved in the "deal" since they are affected by it too (due to the differences in service requirements between Short and Long).

While such a simplistic approach may hold for some cases, the subject is vast and complex and many issues, such as congestion pricing, balking, reneging and many others, remain untouched. This subject is an important and challenging area for future research.

Finally, and maybe above all, experimental studies and publication of actual case studies will advance the area of queue fairness significantly.

5 References

1. B. Avi-Itzhak and H. Levy (2004). On measuring fairness in queues. *Advances of Applied probability*, 36(3):919-936, 2004.
2. N. Bansal and M. Harchol-Balter (2001). Analysis of SRPT scheduling: investigating unfairness, in *Proceedings of ACM Sigmetrics 2001 Conference on Measurement and Modeling of Computer Systems*, 2001, pp. 279-290.
3. M. Bender, S. Chakrabarti and S. Muthukrishnan (1998). Flow and stretch metrics for scheduling continuous job streams, in *Proceedings of the 9th Annual ACMSIAM Symposium on Discrete Algorithms*, 1998, pp. 270-279, San Francisco, CA.
4. J. C.R. Bennett and H. Zhang (1997). Hierarchical Packet Fair Queueing Algorithms. *IEEE/ACM Transactions on Networking*, 5(5):675-689, Oct 1997.
5. T. Bonald and A. Proutiere (2004). On performance bounds for balanced fairness, *Performance Evaluation*, 2004, 55:25-50.
6. E. Brosh, H. Levy and B. Avi-Itzhak (2005). The Effect of Service Time Variability on Queue Fairness, forthcoming.
7. E. G. Coffman, Jr., R. R. Muntz, and H. Trotter (1970). Waiting time distribution for processor-sharing systems. *JACM*, 17:123-130, 1970.
8. R. B. Cooper (1981). *Introduction to Queueing Theory*, Macmillan, 1972. Second Edition, North–Holland (Elsevier), 1981. Also at http://www.cse.fau.edu/~bob/publications/IntroToQueueingTheory_Cooper.pdf.
9. M. F. Chan and J. M. Tien (1981). An Alternative Approach to Police Response, *Wilmington Management of Demand Program*, National Institute of Justice, Washington DC, 1981
10. J. D. Daigle (1992). *Queueing Theory for Telecommunications*, Addison-Wesley, September, 1991. Second Printing, Spring 1992.
11. A. Demers, S. Keshav, and S. Shenker (1990). Analysis and simulation of a fair queueing algorithm. *Internetworking Research and Experience*, 1:3-26, 1990.
12. S. J. Golestani (1994). A Self-clocked Fair Queueing Scheme for Broadband Applications, in *Proceedings of IEEE INFOCOM*, 1994, pp. 636-646.
13. E.S. Gordon (1987). New problems in queues: Social injustice and server production management, MIT, PhD thesis in Operations Research, 1987.
14. A. G. Greenberg and N. Madras (1992). How fair is fair queueing? *JACM*, 3(39):568-598, 1992.
15. D. Gross and C. L. Harris, *Fundamentals of Queueing Theory*, Wiley & Sons, New York, 1974.
16. R. W. Hall (1991). *Queueing Methods for Services and Manufacturing*, Prentice Hall, 1991.
17. M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal (2003). Size-based scheduling to improve web performance, *ACM Transactions on Computer Systems*, 21(2):207-233, May 2003.
18. R. Hassin and M. Haviv (2002). *To Queue or Not to Queue, Equilibrium Behavior in Queueing Systems*, Kluwer Academic Publishers, Boston, 2002.

19. J.M. Jaffe (1981). Bottleneck Flow Control, *IEEE Transactions on Communications*, July 1981, 29(7):954-962.
20. F. P. Kelly (1997). Charging and rate control for elastic traffic, *European Transactions on Telecommunications*, 1997, 8:33-37.
21. F. P. Kelly (1979). *Reversibility and Stochastic Networks.* John Wiley, Chichester, 1979.
22. L. Kleinrock (1975). *Queueing Systems Vol I: Theory*, Wiley, New York, 1975.
23. L. Kleinrock (1976). *Queueing Systems Vol II: Computer Applications*, Wiley, New York, 1976.
24. J. F. C. Kingman (1962). The Effect of Queue Discipline on Waiting Time Variance, *Proc. Camb. Phil. Soc*., 58:163-164, 1962.
25. R. C. Larson (1987). Perspective on queues: Social justice and the psychology of queueing, *Operations Research*. 35(6):895-905, 1987.
26. I. Mann (1969). Queue culture: The waiting line as a social system, *Am. J. Sociol.* 75:340-354, 1969.
27. A. Martin (1983). *Perception and Value Management, Think Proactive*, 8:95-101, 1983.
28. J. T. McEwen, E. F. Connors and M. I. Cohen (1984). *Evaluation of the Differential Police Response Field Test*, Research Management Associates, Inc., Alexandria, Va, 1984.
29. P. Naor (1969). On the Regulation of Queue Size by Levying Tolls, *Econometrica*, 37:15-24, 1969
30. M. Nussbaum (2001). The Enduring Significance of John Rawls, *Chro. High Edu. – The Chronicle Review*, July 20, 2001.
31. C. Palm (1953). Methods of Judging the Annoyance Caused by Congestion, *TELE*, 4:189-108, 1953.
32. A. Parekh (1992). A generalized processor sharing approach to flow control in integrated services networks, MIT, PhD Dissertation, 1992.
33. A. Parekh and R. G. Gallager (1993). A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Trans. Networking*, 1:344-357, June 1993.
34. A. Parekh and R. G. Gallager (1994). A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Trans. Networking*, 2:137-150, 1994.
35. D. Piccard (2005). Outline of an Extended Book Review, *Stanford Encyc Phil*, http://oak.cats.ohiou.edu/~piccard/entropy/rawls.html, Jan. 2005.
36. A. Rafaeli, G. Barron, and K. Haber (2002). The effects of queue structure on attitudes, *Journal of Service Research*, 5(2):125-139, 2002.
37. A. Rafaeli, E. Kedmi, D. Vashdi, and G. Barron (2005). Queues and fairness: A multiple study investigation, Technical Report, Technion – Israel Institute of Technology 2005.
38. J. Rawls (1971, 1999). *A Theory of Justice*, Harvard University Press, 1971, revised edition 1999.
39. D. Raz, H. Levy and B. Avi-Itzhak (2004). A Resource-Allocation Queueing fairness Measure, in *Proceedings of Sigmetrics 2004*. *Performance Evaluation Review*, 32(1):130-141, 2004.

40. D. Raz, H. Levy and B. Avi-Itzhak, (2004a). RAQFM: A Resource Allocation Queueing Fairness Measure, Technical Report RRR-32-2004, RUTCOR, Rutgers University, September 2004.
41. D. Raz, B. Avi-Itzhak, and H. Levy (2004b). Classes, priorities and fairness in queueing systems. RUTCOR Technical Report RRR-21-2004, June 2004.
42. D. Raz, B. Avi-Itzhak and H. Levy (2005). Fairness Considerations in Multi-Server and Multi-Queue Systems. RUTCOR Technical Report RRR-11-2005, February 2005.
43. J. L. Rexford, A. G. Greenberg and F. G. Bonomi (1996). Hardware Efficient Fair Queueing Architectures for High-Speed networks, in *Proc. INFOCOM '96*, pp. 638-646, March 1996.
44. M. H. Rothkopf and P. Rech (1987). Perspectives on Queues: Combining Queues is not Always Beneficial. *Operations Research*, 35:906-909, 1987.
45. W. Sandmann (2005). A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement, in *Proceedings of the 1st Euro NGI Conference on Next Generation Internet Networks Traffic Engineering*, Rome, Italy, pp. 18-20, April 2005.
46. M, Shreelhar and G. Varghese (1996). Efficient Fair Queueing Using deficit Round Robin, *IEEE/ACM Transactions on Networking*, 4(3):375-378, June 1996.
47. *Stanford Encyclopedia of Philosophy* (2005).http://plato.stanford.edu/contents.html.
48. Y.T. Wang and R.J.T Morris (1985). Load sharing in distributed systems, *IEEE Tx. on computers*, C-34(3):204-217. 1985.
49. W. Whitt (1984). The amount of overtaking in a network of queues, *Networks*, 14(3):411-426, 1984.
50. A.Wierman and M. Harchol-Balter (2003). Classifying scheduling policies with respect to unfairness in an M/GI/1, in *Proc. ACM Sigmetrics 2003 Conference on Measurement and Modeling of Computer Systems*, San Diego, CA, pp. 238-249, June 2003.