On Measuring Fairness in Queues

Benjamin Avi-Itzhak
RUTCOR, Rutgers University, New Brunswick, NJ, USA
aviitzha@rutcor.rutgers.edu
Hanoch Levy
School of Computer Science
Tel-Aviv University, Tel-Aviv, Israel
hanoch@cs.tau.ac.il

February 29, 2004

Abstract

The issue of "fairness" is raised frequently in the context of evaluating queueing policies, notably in relation to telecommunications and computer systems where it may be of no lesser importance than the conventional measures of performance. Comparisons of the fairness of various systems and policies are often awkward due to lack of generally accepted definitions and measures for this important property.

The purpose of this work is to propose possible fairness measures enabling to quantitatively measure and compare the level of fairness associated with G/G/R queueing systems. We define and discuss *Order (of service) Fairness* and use an axiomatic approach for developing a measure for it in the G/D/1 case. The measure obtained for the G/D/1 system is then generalized and applied to the G/G/R class of systems.

A practical implication of this work is that, for a wide class of service disciplines, the variance of the waiting time can be used as a yard-stick for comparing fairness levels.

keywords: Queueing, Fairness, Service Order, Queue Disciplines, Customer Scheduling.
Subject classifications: 68M20 (Performance evaluation; queueing; scheduling), 60K25 (Queueing theory).

1 Introduction

The issue of "fairness" is raised frequently in the context of evaluating queueing policies. In many applications, in which an important question is how to operate the queueing system, the two major issues most often considered are those of "efficiency", which is mostly attributed to the mean waiting time in the system, and of "fairness". While the former measure is well understood and generally agreed upon, no common yardsticks exist for measuring "fairness" in queueing systems.

The fairness factor associated to waiting in queues has been recognized in many works and applications. Larson (1988) in his discussion paper on the dis-utility of waiting, recognizes the central role played by 'Social Justice', (which is another name for fairness), and its perception by customers. This is also addressed in Rothkopf and Hech (1987) in their paper discussing perceptions in queues. Aspects of fairness in queues were discussed earlier by quite a 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.

Service scheduling issues in telecommunication systems at large, and with regards to web servers in particular, have become more crucial with the rapid and diverse spreading of these systems use. As a result, the interest in the fairness levels associated with the different schedules has greatly increased and the fairness factor is recently more often referred to in many applied and theoretic publications. However, in these works, no general yardsticks or measures are provided that allow quantitative comparison of fairness across whole classes of schedules (disciplines).

In the telecommunications systems area, fairness has received attention in the context of "flow-fairness" or "stream fairness". In this context a user is associated with a stream of packets and the server (typically a routing unit) aims at scheduling the packets of the various streams in a way that provides proper throughput to the various flows while maintaining some fairness to the sources of the flows. Within this context, Wang and Morris (1985) discuss a type of fairness yardstick in dealing with load sharing in distributed systems. They define fairness of service as "uniformly acceptable performance provided to jobs regardless of the source on which the jobs arrive", and then devise a yardstick for measuring quality of load sharing, named Q-factor. The Q-factor measures the performance, relative to global FIFO (namely FIFO among the customers of all sources), as observed by the customer stream treated worst under the worst possible combination of stream loads. Fairness is thus defined via a measure of extreme treatment by the service provider with respect to customer streams.

Further, within the context of "stream fairness" the literature has flourished on the subject of *Fair Queueing*, which received very much attention in the recent decade. Fair Queueing disciplines (and their variants) aim at the fair scheduling of packet streams within network devices. The departure point for these is that General Processor Sharing (GPS) provides fair treatment of streams and thus policies "close" to GPS, where the packets are served one by one, are fair as well. 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 (1996). 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".

In contrast to streams, for which much "fairness" literature has been written, the fairness of individual job treatment has not been thoroughly discussed in the literature. In individual job treatment we refer to the decisions of scheduling for service individual jobs rather than streams. This relates to the treatment of waiting lines in a very wide range of conventional application areas, including simple systems like banks and supermarkets, at the lower end, up to airports, seaports, and complex production and computer systems at the higher end. In the context of communications networks an example is the treatment of jobs within a Web server. In this application the server receives Web page requests of variable sizes from many users (where typically a request is associated with a user) and must serve them in some order. Fairness among individual requests (jobs) must therefore be dealt with when determining the order of service. A key factor playing a role in individual customers scheduling is that of seniority (measured in time already spent in waiting). which, by contrast, plays a minor role (next to throughput) in stream fairness.

The purpose of this work is to address the issue of customers scheduling fairness. We propose yardsticks that can be used as standards for evaluating the fairness of various queueing systems with one class of customers, and for comparing different disciplines to each other .

As seniority (time elapsed since arrival) plays a major role in job fairness, our proposed measures explicitly account for the seniority of the various customers. Our initial step in developing the fairness measure is to treat the simple case, where service times are a constant, using an axiomatic approach. We base our proposition on several elementary properties that seem to be requisite of a fairness measure and develop the fairness measures based on these properties (axioms).

The paper is structured as follows: Section 2 summarizes the queueing models and classes of disciplines addressed in the body of the paper. In Section 3 we consider ergodic single server queueing systems with constant service times, so that customers are distinguishable only by their arrival times. To develop a fairness measure for such systems we state four elementary properties that seem to be requisite of such a measure, and using an axiomatic approach we construct a measure possessing these properties. It turns out to be proportional to the second moment of the stationary waiting time distribution. In Section 4 we analyze and discuss the suitability of adopting the fairness measure developed for G/D/1 systems to G/G/1 systems. We conclude that the axiomatic properties assumed for constant service times do not hold for stochastic service times and propose, seemingly different, basic properties for the G/G/1 class. Based on the conclusions of Section 4 we develop in Section 5 a fairness measure for the many servers G/G/R class of systems and also show that the elementary axiomatic properties used for G/D/1 are actually a special case. The paper concludes with examples and concluding remarks.

1.1 Alternative Measures

The measure developed in this paper focuses on seniority, since relative seniority is highly important in many queueing facilities and to many customers. For example, seniority is of utmost importance when the service (or the product delivered by the server) is not guaranteed and customers who are pushed back in the queue may find themselves not served at all (e.g. a queue for a finite number of concert tickets).

One may consider an alternative measure that focuses on service times. In particular one may require that in a fair regime the waiting time of each customer be proportional in length to his/her required service. This is achieved, in *expectation*, in symmetric queues (as defined in Kelly (1978)), the LIFO preemptive resume and processor sharing schemes being two examples. While this approach might be appealing from service time perspective, it is problematic from seniority aspects which it practically ignores (thus leading to the result that LIFO preemptive is fair).

An approach that will account for both seniority and service times in its fundamental principles is an issue of an ongoing research.

2 Models

We consider queueing systems consisting either of a single server (Sections 3 and 4) or of multiple servers (Section 5). The service of each customer is rendered by exactly one of the servers (in the case of multiple servers), and upon service completion the customer leaves the system. The order by which customers are selected for service is the service discipline.

We limit this study to work-conserving service disciplines with uninterruptable service and one-customer-at-a-time at each server. Thus systems where service is given in slices (such as Round-Robin) are not considered. When service terminates the vacated server selects a customer from the waiting line, if not empty. Under these conditions, we consider two classes of service disciplines: In the first class, denoted by Φ_1, the selection of customer for service is done without consideration of the service times required by any of the waiting customers. Examples are *FIFO*, *LIFO* and *Random Selection*. In the second class, denoted by Φ_2, the selection of customers for service can be done with consideration of service requirement. An example is *shortest job first*.

Note that all disciplines considered in this work forbid servers idling when waiting lines are non-empty. Also note that in general ($\Phi_2 \supset \Phi_1$). However, in the specific case where service times are constant there is no difference between Φ_1 and Φ_2 disciplines.

The notation to be used will be defined in the paper, where we usually use upper case letters to denote random variables while lower case letters denote their realizations in a particular sample path. For the reader convenience, a glossary of frequent notation is given in the appendix.

3 A Fairness Measure for an Ergodic System with Constant Service Times – G/D/1.

Customers of the same class are distinguished, in general, by their arrival and service times. One has to assume that, in the absence of additional distinguishing variables, a fairness measure will have to involve these two variables. The role of arrival times, which determine the seniority of customers, is clear: violating the seniority of a customer, by serving less senior customers ahead of her, is a source for resentment. As for service times, even in the situation where service times are i.i.d., and known deterministically only after service completion, the realization that customers violating her seniority have long service times is an additional cause for frustration. On the other hand, if the seniority violators happen to have short service requirements her resentment due to her seniority violation may decrease.

In a $G/D/1$ system, where service requirements of all customers are identical and constant, say d time units, the customers are distinguished only by their arrival times. This simplified system provides a tractable starting point for an axiomatic development of a fairness measure. We assume that unfairness frustration of individual customers is attributed to violation of their seniority and that the more pronounced violations result in higher frustration. Hence the most fair service ordering is FIFO, (the 'fairest' queue discipline, in a sense, according to Kingman (1962)), and deviations from it create unfairness, (deviation from "social justice" as used by Larson (1988)). In what follows we use this rationale as a basis for constructing a total fairness measure for the $G/D/1$ system, where the G, as commonly used, means that the arrival process is an arbitrary stationary point process. The term ergodic is used in this paper to imply the assumptions that the process will reach stationarity in the long run and that the long time averages mentioned for that process do exist.

3.1 Axiomatic Properties of Order Fairness in a G/D/1 System

Let C_i, $i = 1, 2, ..$, denote the ith customer (in order of arrival). Let A_i, $i = 1, 2, ...$ be a random variable denoting the arrival time of C_i, and a_i denote the realization of A_i. Let S_i, $i = 1, 2, ..$ be a random variable denoting the service time of C_i, and s_i denote the realization of S_i. Let π denote a realization of service order, (a permutation of the indices $1, 2, \ldots$). Wherever an order π is mentioned in this paper it is assumed to be a feasible order (not all service orders are feasible, e.g., serving first in a busy period the customer who arrives last at the busy period).

Assume the system is ergodic and let $a_1, a_2, \ldots, a_n$ be a realization of arrival times such that at time $t = a_1$ the system's state changes from empty to non-empty and becomes empty again for the first time at $t = a_1 + nb$, i.e. $C_1, C_2, \ldots, C_n$ are served exclusively in one busy period of size n. Assume that only service orders that are feasible under disciplines class Φ_1 are allowed and let $F(\pi)$ be a total fairness measure for $C_1, C_2, \ldots, C_n$ when served in order π. Further assume that C_j and C_k are neighbors in π, i.e., $p_j^\pi = p_k^\pi + 1$ and let π' be the feasible order resulting from interchanging C_j and C_k in π. Then the change in fairness $f_{j,k,\pi}$ due to this neighbors interchange is given by:

$$f_{j,k,\pi} = F(\pi') - F(\pi) \ . \tag{1}$$

We axiomatically assume the following four properties for $f_{j,k,\pi}$:

(P.1) – Monotonicity in seniority under neighbors interchange: For neighbors C_j and C_k in π $f_{j,k,\pi}$ is monotone, strictly increasing, in their seniority difference $a_k - a_j$. Note the equality of simultaneously arriving customers: $f_{j,k,\pi} = 0$ if $a_j = a_k$.

(P.2) – Reversibility of neighbors interchange (symmetry): If π' is obtained by a neighbors interchange of C_j and C_k in π, then $f_{k,j,\pi'} = -f_{j,k,\pi}$.

(P.3) – Independence on position and time: The fairness change $f_{j,k,\pi}$ is invariant with respect to the positions of the neighbors C_j and C_k in π and to the time at which the change occurs.

(P.4) – Fairness change unaffected by customers not interchanged: $f_{j,k,\pi}$ is functionally independent of the arrival times a_i and positions p_i^π of customers other than C_j and C_k.

The rationale behind (P.1) was given in the beginning of this section. (P.2) ensures that total fairness returns to its original level when a neighbors' interchange is reversed. (P.3) and (P.4) provide that $f_{j,k,\pi}$ is the same for all π where C_j and C_k are interchangeable neighbors, and when combined with (P.1) we have

$$f_{j,k,\pi} = f_{j,k} = g(a_j - a_k) \ , \tag{2}$$

where g, the change function corresponding to F, is monotone increasing. A seemingly appealing rationale behind (P.4) is that interchanging neighbors affects only these two customers and therefore the change in total fairness should be determined wholly by the parameters of the interchanged customers only. A possible generalization of (P.4) is to require that it would hold for non-neighbor interchanges as well, namely:

(P.4G) – **Change due to non-neighbor interchange unaffected by customers not interchanged:** The fairness change introduced by the interchange of the non-neighboring customers C_j and C_k is functionally independent of the arrival times a_i and positions p_i^π of customers $C_i \neq C_j$ or C_k.

Note however, that in general the rationale behind (P.4) and (P.4G) may not necessarily coincide with the notion of fairness which is mostly reflective of a sense of equity in treatment received. Assuming that quality is represented by the waiting time which measures the efficiency of the system in terms of quality of service delivered, then the differences in individual waiting times, when compared to one another, measure the equity. The interchange of two customers with identical service times does not affect the waiting times of any other customers. Nevertheless, it does affect the differences between the individual waiting times of not only the two interchanged customers, and thereby it does affect the equity of all the customers of the system. This issue is dealt with in Sections 4 and 5 of this paper.

The following three lemmas result directly from the basic properties, (P.1) – (P.4):

Lemma 1 (Insensitivity to Particular Sequence of Interchanges): *If π_1 is transformed into π_2 by a sequence of neighbors interchanges, the fairness difference between π_1 and π_2 is indifferent to the particular sequence.*

Proof. Consider two arbitrary customers C_j and C_k and let us trace the process of transforming π_1 into π_2: During this process C_j and C_k may possibly interchange more than once. Let $A_{j,k}$ be the number of such interchanges in which C_k is moved ahead of C_j and let $A_{k,j}$ be the number of such interchanges in which C_j is moved ahead of C_k. Now, the difference between these two numbers, $A_{j,k} - A_{k,j}$ must be

$$A_{j,k} - A_{k,j} = \begin{cases} -1 & \text{if } p_j^{\pi_1} > p_k^{\pi_1} \text{ and } p_j^{\pi_2} < p_k^{\pi_2} \\ 1 & \text{if } p_j^{\pi_1} < p_k^{\pi_1} \text{ and } p_j^{\pi_2} > p_k^{\pi_2} \\ 0 & \text{otherwise (no change in relative order of } C_j \text{ and } C_k) \end{cases} \tag{3}$$

which is independent of any particular sequence. From (P.2)–(P.4) we have that the change in fairness due to neighbors interchanges of C_j and C_k equals to $(A_{j,k} - A_{k,j})g(a_j - a_k)$. Summing over all (j,k) pairs, $k \neq j$, we obtain the total change in fairness when transforming π_1 to π_2 as

$$F(\pi_2) - F(\pi_1) = \frac{1}{2}\sum_{j,k}(A_{j,k} - A_{k,j})g(a_j - a_k) \ , \tag{4}$$

and the lemma is shown to be true. □

Lemma 1 implies that the selected change function g defines the particular total fairness measure $F(\pi)$ up to an additive constant.

Lemma 2 (Fairness of FIFO): *The FIFO service order is the most fair; its degree of fairness is the highest any ordering can have.*

The proof is immediate by noting that any order π can be obtained from π_{FIFO} by a sequence of neighbors interchanges in which each interchange results in advancing the service of a customer on the account of a more senior customer (that gets pushed back). Due to (P.1), each of these steps reduces the fairness measure of the system, and thus the final fairness measure of π is lower than that of π_{FIFO}.

Lemma 3 (Time-shift invariance of fairness measure): *If arrival times are shifted by a constant, i.e., the sequence $a_1, a_2, \ldots$ is replaced by $a_1 + a, a_2 + a, \ldots$, for some constant a, the fairness measure is not affected.*

This is a consequence of $f_{j,k,\pi}$ being a function of the seniority difference $a_k - a_j$ only (P.1) and independent of the parameters of other customers (P.4).

3.2 A Proposed Measure for Total Order Fairness

Let δ_i^π denote the realized order displacement of C_i in π, that is

$$\delta_i^\pi = p_i^\pi - i, \tag{5}$$

where p_i^π is the position of C_i in π. Let d_i^π denote the waiting displacement of C_i in a realization with order π. This is the change in waiting time due to the order displacement δ_i^π (e.g. if service times are constant b in a $G/D/1$ system then $d_i^\pi = b\delta_i^\pi$). The following functional form is proposed for the total fairness $F(\pi)$, and total unfairness $\mathrm{UF}(\pi)$, when $C_1, C_2, \ldots, C_n$ are served in order π:

$$F(\pi) = -\,\mathrm{UF}(\pi) = c\sum_{i=1}^{n} a_i d_i^\pi + \alpha\ , \tag{6}$$

where $c > 0$ is an arbitrary constant and α is an arbitrary function of $a_1, a_2, \ldots, a_n$ and n. (Note that the value of α is invariant for all π feasible for a given set $a_1, a_2, ..., a_n$ and n).

From the definition of d_i^π we have

$$F(\pi) = cb\sum_{i=1}^{n} a_i \delta_i^\pi + \alpha\ . \tag{7}$$

We note that when $\alpha = 0$, this functional form pegs the fairness and unfairness values of π_{FIFO} to 0 and that $F(\pi) \leq 0$ for all feasible π.

Theorem 1 *The fairness measure $F(\pi)$ possesses the axiomatic properties (P.1)-(P.4).*

Proof. To prove (P.1), consider a policy π and construct policy π' from π by interchanging C_k and C_j (whose positions in π are $p_j^\pi = p_k^\pi + 1$). Assuming π' is feasible, the resulting fairness change is

$$f_{j,k,\pi} = F(\pi') - F(\pi) = c\sum_{i=1}^{n} a_i b \delta_i^{\pi'} - c\sum_{i=1}^{n} a_i b \delta_i^{\pi}\ . \tag{8}$$

Substitution of 5 yields

$$f_{j,k,\pi} = cb \sum_{i=1}^{n} a_i (p_i^{\pi'} - p_i^{\pi}) \ . \tag{9}$$

But by the definition of π and π' we have $p_i^{\pi} = p_i^{\pi'}$ for all $i \neq j$ and $i \neq k$, $p_k^{\pi'} - p_k^{\pi} = +1$ and $p_j^{\pi'} - p_j^{\pi} = -1$, hence

$$f_{j,k,\pi} = cb(a_k - a_j) \ . \tag{10}$$

This is a monotone increasing and symmetric function of the seniority difference of the two interchanged customers only. The proof for (P.2)–(P.4) is immediate. □

Clearly, it is possible to come up with any number of different functional forms of g producing different fairness measures, all of which satisfy the four basic required properties (P.1)–(P.4). We note, however, that the measure $F(\pi)$ as defined in (6), satisfies (P.4) for non-neighbors interchanges too, namely for C_j and C_k which are not necessarily neighbors. This means that it satisfies (P.4G) as well, and as such it is unique.

The uniqueness of the measure $F(\pi)$ is shown next.

Theorem 2 *A fairness function satisfying (P.1)-(P.4) will satisfy (P.4G) if and only if it is of the form $F(\pi)$ given in (6).*

Proof. To show that $F(\pi)$ satisfies (P.4G), suppose we interchange C_j and C_k whose positions in a given order π are $\ldots C_j, C_{j_1}, C_{j_2}, \ldots C_{j_m}, C_k \ldots$, i.e., they are separated by m customers, $C_{j_1}, C_{j_2}, \ldots C_{j_m}$. This can be achieved by first moving C_k $(m+1)$ positions to the left and then moving C_j m positions to the right. The resulting order is denoted by π' and assumed feasible. The change in fairness is given by

$$F(\pi') - F(\pi) = g(a_k - a_j) + \sum_{i=1}^{m} [g(a_{j_i} - a_j) + g(a_k - a_{j_i})]. \tag{11}$$

Substituting the linear form of the change function from (10) we have

$$g(a_{j_i} - a_j) + g(a_k - a_{j_i}) = cb(a_k - a_j) \ , \tag{12}$$

and therefore

$$F(\pi') - F(\pi) = (m+1)cb(a_k - a_j)\ , \tag{13}$$

which satisfies (P.4G).

To show that $F(\pi)$ is the only function satisfying (P.4G) assume that another function, $\phi(\pi)$, also satisfies this property and let g be its corresponding change function. Next take the order π_{FIFO} and interchange the positions of C_j and C_{j+2} (assuming that this interchange is feasible), by moving C_j two positions to the right and then moving C_{j+1} one position to the left. The resulting change in the fairness measure is then equal to

$$g(a_{j+1} - a_j) + g(a_{j+2} - a_j) + g(a_{j+2} - a_{j+1})\ . \tag{14}$$

For (P.4G) to hold, this expression must be a function of a_j and a_{j+2} only, namely there exists a function h such that

$$g(a_{j+1} - a_j) + g(a_{j+2} - a_{j+1}) = h(a_j, a_{j+2})\ . \tag{15}$$

For any fixed given values of a_j and a_{j+2} the right hand side is constant when a_{j+1} is varied in the interval $[a_j, a_{j+2}]$. Noting that $g(0) = 0$ and setting $a_{j+1} = a_j$ in (15) we obtain

$$h(a_j, a_{j+2}) = g(a_{j+2} - a_j)\ . \tag{16}$$

Defining $x = a_{j+1} - a_j$ and $y = a_{j+2} - a_{j+1}$ we obtain from (15) and (13) that

$$g(x) + g(y) = g(x+y)\ . \tag{17}$$

Thus, by the definition of a linear function, $g()$ is a linear function of the difference in seniority and therefore $\phi(\pi)$ is identical to $F(\pi)$, up to a multiplicative constant. $\square$

For simplicity we will assume in the remainder of this chapter and the next one that $\alpha = 0$. For consistency we further assume that c is the same positive constant for all realizations of a busy period.

3.3 Expected Order Fairness Per Customer

For the realization of the busy period with $C_1, C_2, \ldots, C_n$ the order of service, π, is determined by the particular queue discipline applied. E.g. if the waiting line discipline is FIFO then $\pi = (1, 2, \ldots, n)$ always, while for disciplines such as LIFO, Random Selection for Service and all other disciplines in Φ_1, the order π will depend on the discipline and the particular realization of arrival times.

For a given discipline, $\varphi \in \Phi_1$, let the random variable F_i be the total fairness of the ith busy period, resulting from (6), and let $\overline{F}$ be the expected fairness per customer.

Assume that the system is in equilibrium. Let Δ_i be a random variable denoting the order displacement of C_i for the given discipline (with corresponding realization δ_i^{π}). Let D_i be a random variable denoting the waiting displacement of C_i for the given discipline

$$D_i = W_i - W_i^*, \tag{18}$$

where W_i^* is the waiting time for FIFO, unless otherwise defined, and W_i is the waiting time for the given discipline.

Then, for $\alpha = 0$ and c a positive constant, we have

$$\overline{F} = \lim_{k\to\infty} \frac{\sum_{i=1}^{k} F_i}{\sum_{i=1}^{k} N_i} = c \lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} A_j D_j = cb \lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} A_j \Delta_j \ . \tag{19}$$

where N_i is the number served in the i-th busy period. In a $GI/D/1$ queue

$$\overline{F} = \frac{E(F)}{E(N)} \ , \tag{20}$$

where F_i are i.i.d as F and N_i are i.i.d. as N. Note that the expected fairness per customer for the FIFO discipline is $\overline{F}^* = 0$.

Theorem 3 *Let W and W^* be the $G/D/1$ equilibrium waiting times under our given discipline and under FIFO, respectively. Then*

$$\overline{F} = \frac{c}{2}[Var(W^*) - Var(W)] \tag{21}$$

Proof. Denote the waiting time of C_j under our given discipline by W_j. Then

$$W_j = E_j - A_j \ , \qquad j = 1, 2, , \ldots , \tag{22}$$

where E_j is the time C_j enters service. Similarly

$$W_j^* = E_j^* - A_j, \qquad j = 1, 2, \ldots , \tag{23}$$

where W_j^* and E_j^* are the waiting and entry-into-service times of C_j, respectively, under the FIFO discipline.

From (22) and (23) we obtain

$$W_j^{*^2} - W_j^2 = 2A_j(E_j - E_j^*) + (E_j^{*^2} - E_j^2) \ . \tag{24}$$

Substituting

$$D_j = E_j - E_j^* \ , \tag{25}$$

results in

$$W_j^{*^2} - W_j^2 = 2A_jD_j + (E_j^{*^2} - E_j^2) \tag{26}$$

Let $B_i = (j_i, j_i + 1, j_i + 2, j_1 + N_i - 1)$ be the set of indices of the N_i customers served in a busy period i. Then because of the constant service times

$$\sum_{j \in B_i} E_j^{*^2} - \sum_{j \in B_i} E_j^2 = 0 \ , \qquad i = 1, 2, \ldots , \tag{27}$$

and thus

$$\sum_{j \in B_i} W_j^{*^2} - \sum_{j \in B_i} W_j^2 = 2 \sum_{j \in B_i} A_j D_j = \frac{2}{c} F_i \ . \tag{28}$$

By taking expectations on both sides, or by taking long-run averages we obtain

$$E(W^{*^2}) - E(W^2) = \frac{2}{c} \overline{F} \ . \tag{29}$$

$E(W)$ is invariant $\forall \varphi \in \Phi$, therefore

$$\overline{F} = \frac{c}{2} [\mathrm{Var} W^* - \mathrm{Var}(W)] \tag{30}$$

or

$$\overline{\mathrm{UF}} = -\overline{F} = \frac{c}{2}[\mathrm{Var}(W) - \mathrm{Var}(W^*)] \tag{31}$$

$\square$

The variance of the waiting time may thus serve as a measure of unfairness, the larger the variance is, the greater is the unfairness of the discipline. It follows then that fairness is an expression of equity of waiting times. Among the Φ_1 class of disciplines, FIFO is the most equitable and therefore $\overline{F} \leq 0$ for all $\varphi \in \Phi_1$. For other classes of disciplines, where FIFO is not necessarily the most equitable, the form of $\overline{F}$ may have to be modified.

4 Order Fairness in an Ergodic $G/G/1$ System

Here we assume that the service times $S_i, i = 1, 2, \ldots,$ are independent of the arrival times and are i.i.d. as S with mean b. As discussed in the previous chapter, customers in this system are distinguished by both their arrival time and their service time. Also, unlike the $G/D/1$ system where $\Phi_1 = \Phi_2$, here $\Phi_1 \subset \Phi_2$ and within Φ_2 the FIFO discipline may not be the most fair.

We first start with assuming class Φ_1 disciplines and adopting the previous chapter results by defining the total fairness of the first busy period as

$$F_1 = c \sum_{j=1}^{N_1} A_j D_j \tag{32}$$

where A_j and D_j $j = 1, 2, \ldots, N_1$ are the arrival time of the j-th arriving customer of the busy period and her waiting displacement respectively, and N_1 is the busy period size. (We use the first busy period for simplicity of notation. The analysis applies to an arbitrary busy period.) Substituting $D_j = W_j - W_j^*$ we have

$$F_1 = c \sum_{j=1}^{N_1} A_j (W_j - W_j^*) \ . \tag{33}$$

To show that Theorem 3 holds also for the $G/G/1$ system we need to prove the following Lemma:

Lemma 4 *Consider a $G/G/1$ busy period of size N_1. For $N_1 \geq i$, $i = 1, 2, \ldots$, we denote by $\mathcal{E}_i$ the i-th entry time into service, and by X_i the number of waiting customers immediately after $\mathcal{E}_i$. (Note that, $\mathcal{E}_i$ is not necessarily the same as E_i). Then, the distributions of $\mathcal{E}_i$ and X_i are the same for all disciplines $\varphi \in \Phi_1$.*

Proof. (By induction): By definition $\mathcal{E}_1 = A_1$ and $X_1 = 0\ \forall \varphi \in \Phi_1$. Let $S^{(k)}$ be the service time of the k-th customer to enter service in the busy period, given $N_1 > k$. Then

$$P(S^{(1)} \leq s) = P(S \leq s \mid S < A_2 - A_1) \quad \forall \varphi \in \Phi_1 \ . \tag{34}$$

Thus

$$\mathcal{E}_2 = \mathcal{E}_1 + S^{(1)} \tag{35}$$

obeys the same probability law for all $\varphi \in \Phi_1$. At time $\mathcal{E}_2$ one of the arrivals in $(\mathcal{E}_1, \mathcal{E}_2]$ is selected for service and X_2 is then distributed as

$$P(X_2 \geq x) = P(\mathcal{E}_2 \geq A_{x+1}) \ , \qquad x = 0, 1, \ldots \tag{36}$$

which is the same $\forall \varphi \in \Phi_1$. Assume now that the Lemma is true for $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k$ and X_1, $X_2, \ldots, X_k$. At time $\mathcal{E}_k$ a customer with service time distributed as S enters into service under all $\varphi \in \Phi_1$. Therefore

$$\mathcal{E}_{k+1} = \mathcal{E}_k + S, \text{ in distribution if } X_k > 0 \tag{37}$$

and

$$\mathcal{E}_{k+1} = \mathcal{E}_k + S^{(k)} \ , \text{ if } X_k = 0 \ , \tag{38}$$

where

$$P(S^{(k)} \leq s) = P(S \leq s \mid S > A_{k+1} - \mathcal{E}_k) \ . \tag{39}$$

Hence the probability law of $\mathcal{E}_{k+1}$ is the same $\forall \varphi \in \Phi_1$. Also

$$P(X_{k+1} \geq x) = P(\mathcal{E}_{k+1} \geq A_{x+k+2}) \quad \forall \varphi \in \Phi_1 \ . \tag{40}$$

□

Corollary 1 *$E(\sum_{i=1}^{N}(E_i)^n)$, $n \geq 1$, is invariable* $\forall\varphi \in \Phi_1$.

The proof follows immediately from Lemma 4.

Corollary 2 *Theorem 3 holds also for the $G/G/1$ case* $\forall\varphi \in \Phi_1$.

Proof. The analysis in the proof of Theorem 3 applies up to Equation (27) which does not hold for $G/G/1$. Therefore, instead of (28) we have

$$\sum_{j\in B_i} W_j^{*^2} - \sum_{j\in B_i} W_j^2 = \frac{2}{c}F_i + \sum_{j\in B_i}(E_j^{*^2} - E_j^2) \ . \tag{41}$$

But from Corollary 1 we have that when taking expectations the second term on the right vanishes $\forall\varphi \in \Phi_1$, resulting in (29) and (31). □

Although the measure obtained for $\overline{F}$ turns out to be the same as in the constant service times case, not surprisingly it does not satisfy the axiomatic properties laid out in chapter 3 for the special case of constant service times. In fact the form we obtain from (33) for the fairness change due to neighbor interchange is

$$f_{j,k,\pi} = c(a_k s_j - a_j s_k) \tag{42}$$

which is a function of the two customers' distinguishing variables. This form satisfies (P.1)-(P.4G) only if all service times are identical, equaling b.

There is a need, thus, to formulate the basic properties of the fairness measure to encompass a wider class of systems and disciplines. The Φ_1 disciplines class is particularly appealing since efficiency, when measured by average waiting time, is the same for all $\varphi \in \Phi_1$, which eliminates the issue of trade-off between efficiency and fairness when comparing different disciplines within this class. Nevertheless, class Φ_1 is too restrictive and in what follows we expand our analysis to class Φ_2 disciplines.

Following the conclusions of chapter 3 we use two premises in the proceeding: (i) unfairness is mainly a result of the inequity in the treatment customers receive, and (ii) unfairness is a group property. Each member of the served group compares the treatment it receives to the treatment received by each other member of the group. The group's overall unfairness per customer is then calculated using all the pair-wise comparisons.

To demonstrate a pair-wise comparison consider the following situation at a supermarket counter: Mr. Short arrives at the counter holding only one item. Upon arrival, Short finds Ms. Long ahead of him with two carts loaded with items. Recognizing the "unfair" situation, in which Short will have to wait for the long processing of Long (while alternatively Long would have to wait very little for the processing of Short), he would ask Long whether interchange of their places is possible. While this proposition may sound as a "fair" one in many situations, especially if Long has also just arrived at the queue, this is not always the case. For instance, if Long arrived to the queue a long time ago and had to wait for the processing of many other customers ahead of her, she would feel the interchange would be unjust (claiming, e.g., "Had I not had to wait so long, I would have been out of here long before Short arrived!"). Thus Long assesses the fairness of the requested interchange on a comparison of her total waiting time to that of Short, rather than considering just her future waiting time.

To formalize the pair-wise comparison suppose that customers C_j and C_k arrived at the system at times $a_j < a_k$, requiring service times s_j and s_k, are considered for service at time $t > a_k$. The choice faced by the server is serving C_j first or C_k first. A dilemma shows up when $s_j > s_k$, in which case C_j may be preferred due to its earlier arrival, while C_k may be preferred due to its shorter service time. Suppose C_j and C_k are the only customers of the system under consideration, and their waiting time difference is $|w_j - w_k|$, where w_j and w_k are their waiting times, respectively. From the preceding discussion and the two premises (i) and (ii) it follows that the unfairness of this group of two customers should be presented by some monotone increasing function of $|w_j - w_k|$, say $h(x)$, $x \geq 0$. To determine if the order (j, k) is better than (k, j) we note that

$$h(|w_j - w_k|) = \begin{cases} h(|a_k - a_j - s_j|) & \text{if } (j, k) \\ h(|a_k - a_j + s_k|) & \text{if } (k, j) \ . \end{cases} \tag{43}$$

It is easily shown that order (j, k) is better if

$$a_k - a_j > \frac{1}{2}(s_j - s_k) \ . \tag{44}$$

Note that in this group of two customers the unfairness per customer is $\frac{1}{2}h(|w_j - w_k|)$.

If C_j and C_k are not the only system's customers, and are not necessarily neighbors, relation (44) can still be shown to be the criterion for minimizing the pair-wise unfairness, $h(|w_j - w_k|)$. However, interchanging the order of service of two customers, whether neighbors or not, is bound, in general, to change the pairwise equity of other pairs served in the system.

5 Order Fairness in an Ergodic $G/G/R$ System

To define the unfairness of a many server system, $G/G/R$, we will assume that the unfairness per customer is taken as the average pair-wise unfairness (which is per two customers), divided by two. Specifically, suppose the system stops accepting customers after the n-th arrival and let (a_i, s_i), $i = 1, 2, \ldots, n$ be a realization of the arrival and service times respectively of $C_i, i = 1, 2, \ldots$. For a given service order π the unfairness per customer is given by

$$\overline{\mathbb{UF}}(\pi, n) = \frac{1}{2n^2} \sum_{k=1}^{n} \sum_{j=1}^{n} h(|w_j - w_k|) \ , \tag{45}$$

where w_i, $i = 1, 2, \ldots, n$, are the waiting times realized under order π.

The total unfairness of the group is

$$\mathbb{UF}(\pi, n) = n\overline{\mathbb{UF}}(\pi, n) = \frac{1}{2n} \sum_{k=1}^{n} \sum_{j=1}^{n} h(|w_j - w_k|) \ , \tag{46}$$

and the ergodic expected unfairness per customer is

$$\overline{\mathbb{UF}} = \lim_{n\to\infty} \frac{1}{2n^2} \sum_{k=1}^{n} \sum_{j=1}^{n} h(|W_j - W_k|) \tag{47}$$

Theorem 4 *Let W and W^* be the equilibrium waiting times under disciplines φ and φ^* in Φ_2 respectively, where $Var(W^*) \leq Var(W) \ \ \forall \varphi \in \Phi_2$. If $h(x) = c_0 x^2 + c_1$, where $c_0 = c/2$ and $c_1 = -cVar(W^*)$ for some arbitrary $c > 0$, then*

$$\overline{\mathbb{UF}} = \frac{c}{2}[Var(W) - Var(W^*)] \ , \tag{48}$$

which is the same as in Theorem 3.

Proof. To prove the theorem we will first prove the following Lemma:

Lemma 5 *Let W_i be the waiting time of C_i in an ergodic queuing system, then*

$$\lim_{n\to\infty} \frac{1}{2n^2} \sum_{k=1}^{n} \sum_{j=1}^{n} (W_j - W_k)^2 = Var(W) \ , \tag{49}$$

where W is the equilibrium waiting time.

Proof of Lemma: By expanding the squared terms in (49) we obtain, after some re-arrangement:

$$\frac{1}{2n^2} \sum_{k=1}^{n} \sum_{j=1}^{n} (W_j - W_k)^2 = \frac{1}{n} \sum_{i=1}^{n} W_i^2 - \frac{1}{n^2} \Big(\sum_{i=1}^{n} W_i \Big)^2 \ . \tag{50}$$

Taking the limit as $n \to \infty$ completes the proof of the Lemma.

For

$$h(|W_j - W_k|) = c_0 (W_j - W_k)^2 + c_1 \tag{51}$$

we obtain from (47) and Lemma 5

$$\overline{\mathbb{UF}} = c_0 \lim_{n\to\infty} \frac{1}{2n^2} \sum_{k=1}^{n} \sum_{j=1}^{n} (W_j - W_k)^2 + c_1 = c_0 \mathrm{Var}(W) + c_1 \tag{52}$$

substituting $c_0 = c/2$ and $c_1 = -cVar(W^*)$ yields (48). $\square$

We note that by selecting $c_0 = 1$ and $c_1 = 0$ we obtain $\overline{\mathbb{UF}} = Var(W)$.

Theorem 5 *If the service times of all customers are identical, equaling to a constant b, and $R = 1$, namely $G/D/1$, the unfairness measure $\mathbb{UF}(\pi, n)$ satisfies all the axiomatic properties (P.4G included) of Section 3 if and only if $h(x) = c_0 x^2 + c_1$, where $c_0 > 0$ and c_1 are arbitrary constants.*

Proof. Substituting $h(|\omega_j - \omega_k|) = c_0(\omega_j - \omega_k)^2 + c_1$ in (46) we have

$$\overline{\mathbb{UF}}(\pi, n) = \frac{c_0}{2n} \sum_{k=1}^{n} \sum_{j=1}^{n} (\omega_j - \omega_k)^2 + \frac{nc_1}{2} \ . \tag{53}$$

Expanding the squared terms and then summing we obtain

$$\overline{\mathbb{UF}}(\pi, n) = c_0 \sum_{j=1}^{n} \omega_j^2 - \frac{c_0}{n} \Big(\sum_{j=1}^{n} \omega_j \Big)^2 + \frac{nc_1}{2} \ . \tag{54}$$

By definition $\omega_j = e_j - a_j$ and $e_j = e_j^* + d_j$ where e_j and e_j^* are the values of E_j and E_j^*, respectively, in the given realization. Therefore

$$\omega_j^2 = e_j^2 - 2a_j e_j^* + a_j^2 - 2a_j d_j \ . \tag{55}$$

But in the $G/D/1$ case $\sum_{j=1}^{n} e_j^2$ is invariant for all feasible π and therefore summing (55) yields

$$\sum_{j=1}^{n} \omega_j^2 = \sum_{j=1}^{n} {\omega_j^*}^2 - 2 \sum_{j=1}^{n} a_j d_j \ . \tag{56}$$

Also in the G/D/1 case $\sum_{j=1}^{n} \omega_j$ is invariant for all feasible π. Upon substitution in (54) we get

$$\overline{\mathbb{UF}}(\pi, n) = -2c_0 \sum_{j=1}^{n} a_j d_j + \Big[\sum_{j=1}^{n} {\omega_j^*}^2 - \frac{1}{n} \Big(\sum \omega_j^* \Big)^2 + \frac{nc_1}{2} \Big] \ . \tag{57}$$

The term in the square brackets on the right is the same for all feasible π, thereby

$$F(\pi, n) = -\overline{\mathbb{UF}}(\pi, n) = c\sum_{j=1}^{n} a_j d_j + \alpha \tag{58}$$

which is identical to (6). $\square$

6 Examples

The average pairwise unfairness measure proposed in Section 5 enables comparing the fairness of given disciplines in Φ_2. The question of how to develop high fairness disciplines remains open. In the following we propose a simple greedy heuristic algorithm, based on the analysis of Sections 4 and 5, for implementing a high fairness discipline.

Assume that the pairwise unfairness function $h(x)$ is monotone increasing in $|w_j - w_k|$, then the condition for ordering any two customers, C_j and C_k, as provided in (44), is independent both of the time at which it is done and of the other customers present at the queue at that time. Therefore, we propose to use a simple *index rule* in which the customers are sorted according to *increasing order* of the index:

$$\frac{1}{2}s_i + a_i. \tag{59}$$

Such sorting can be done at any moment without being reversed later. Moreover, since the prioritization rule is independent of the number of customers served between C_j and C_k, the decision taken regarding the relative order of these two customers will not be affected or reversed later due to future arrivals.

The algorithm keeps the customers, which are waiting for service, in a queue which is sorted according to the index $\frac{1}{2}s_i + a_i$. When a new customer arrives its index value is evaluated and the customer is placed in the queue in the appropriate place. When a service is completed the next customer to be served is taken from the head of the queue.

We note that this scheduling algorithm is identical to the FIFO discipline when all service times are identical. Nevertheless, as expected from the discussion following relation (59), the greedy algorithm does not necessarily maximize the all-pairs fairness of a queueing system, as shown in the following example.

Consider a system with two types of customers: type-0 have zero service time and type-1 have service time whose length is one unit. The arrivals are characterized by a bulk arrival process in which every bulk consists of three type-1 customers and one type-0 customer. The inter-arrival times are greater than 3 units so that the likelihood of one bulk arriving while another bulk is in service is zero. Thus, the four members of a bulk form an independent busy period. The issue of scheduling, therefore, boils down to the question of how to schedule the four members of the bulk in order to maximize the fairness measure.

The Greedy algorithm will schedule the service of the type-0 customer first, since a pairwise comparison of a type-0 and a type-1 customer yields priority to the type-0 in this case. The expected unfairness per customer, denoted by $\overline{\mathbb{UF}}_1$, is in this case

$$\overline{\mathbb{UF}}_1 = \frac{1}{32}\left(h(0) + 3h(1) + 2h(2)\right). \tag{60}$$

An alternative approach is to schedule the service of type-0 customer, in which case the expected unfairness per customer, $\overline{\mathbb{UF}}_2$, is

$$\overline{\mathbb{UF}}_2 = \frac{1}{32}\left(h(0) + 4h(1) + h(2)\right). \tag{61}$$

This service ordering results in lower unfairness per customer

$$\overline{\mathbb{UF}}_1 - \overline{\mathbb{UF}}_2 = \frac{1}{32}\left(h(2) - h(1)\right) \geq 0, \tag{62}$$

where the inequality is strict if h is strictly increasing.

A typical elementary example that demonstrates the issues of fairness and efficiency is that of scheduling customers of two types in a single service system. Consider a single server with a single queue. Customers of two types, type-1 and type-2, arrive at Poisson rates λ_1 and λ_2 and require constant service times s_1 and s_2, respectively (assume $s_1 < s_2$). An important question in this setting is what service policy should be used in order to improve the system's performance.

It is well known (see Cox and Smith (1961)), that if the objective is to minimize the mean customer delay then the optimal policy is to prioritize the type-1 customers (shorter jobs first). Other policies, like FIFO will lead to larger mean delay.

There is a question, then, how different policies compare in fairness and not only in efficiency. The argument that giving high priority to type-1 customers (shorter jobs first) is the "fair policy", (since placing a short job ahead of a long job creates less harm than the reverse), disregards the seniority difference between the customers, which is one of the major determinants of fairness. This is demonstrated in Figures 1 and 2, in which we depict the unfairness (expressed as the standard deviation of the waiting time) as a function of the traffic load, for several frequently mentioned disciplines and for the pairwise heuristic scheme. The results were derived from a one million customers simulation.

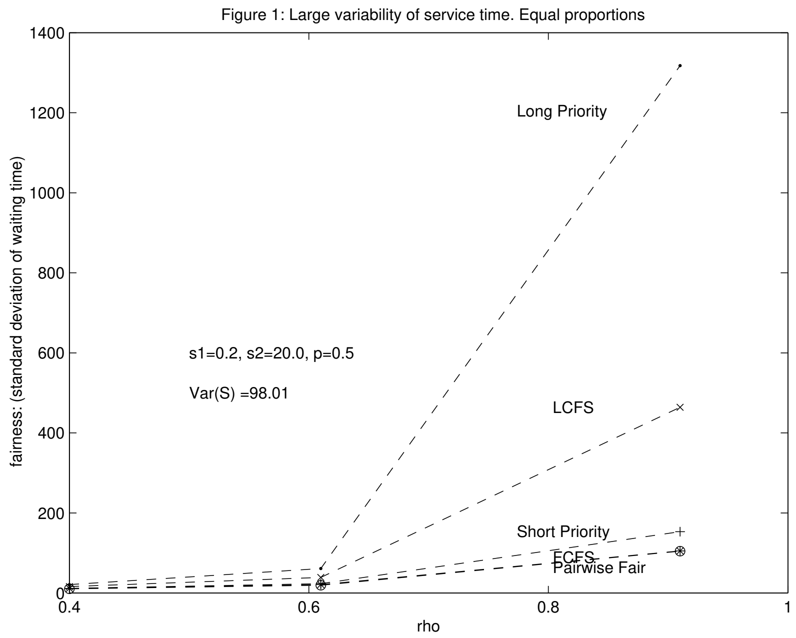

Figure 1: Unfairness of various policies (equal proportions of customer types)

In this figure $p = \lambda_1/(\lambda_1 + \lambda_2) = 0.5$, and there is a large variability in service requirements, $s_2 = 100 s_1$. The most unfair scheme is giving the priority to long service customers. This is also the most inefficient scheme. However, giving the priority to the short service customers, which is most efficient, is less fair than both FIFO and the pairwise dynamic scheme. Changing the customers mix to $p = 0.99$, see Figure 2, results in the priority-to-short-service-customers becoming the most fair scheme. In this case there is no conflict between efficiency and fairness.

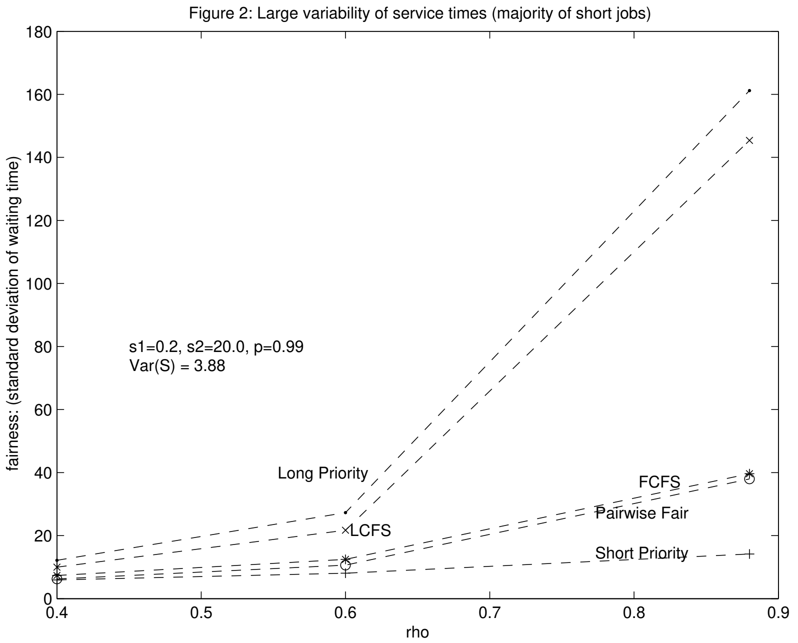

Figure 2: Unfairness of various policies (majority of short jobs)

7 Concluding Remarks

The objective of this work has been to study the issue of fairness in queues and to develop a quantitative measure that can be used to measure fairness in queues. We focused on the issue of customer seniority which is crucial in many queueing systems and used an axiomatic approach to develop a fairness measure that is based on this notion.

Little has been studied on the subject of fairness in queues. More research is needed as to better understand this subject and to build an intuition of how to quantify fairness in queues. In particular the intricate inter-play between seniority and service times, and their effect on queue fairness, need to be better understood. This issue is the subject of an ongoing research.

Acknowledgment: We would like to thank R. McLean for early discussions on the subject of fairness. We are grateful to Lars Nordmann for assistance in the simulation program. We are thankful to the anonymous referees whose suggestions significantly enhanced the presentation of the paper.

References

[1] J.C.R Bennet and H. Zhang (1996), "WF2Q: Worst case fair weighted fair queueing", IEEE INFOCOM'96. San Francisco, March 1996, 120-128.

[2] D. R. Cox and W.L. Smith (1961), "Queues," John Wiley & Sons, New York 1961.

[3] A. Demers, S. Keshav and S. Shenker (1990), "Analysis and simulation of a fair queueing algorithm," *Internetworking: research and experience*, Vol 1, 3-26.

[4] S. J. Golestani (1994), "A self Clocked Fair Queueing Scheme for Broadband Applications," IEEE INFOCOM '94, Toronto, Ontario, June 1994, 636-646.

[5] A. G. Greenberg and N. Madras (1992), "How Fair is Fair Queueing?," *Journal of the ACM*, Vol.3, No. 39, 568-598.

[6] F.P. Kelly (1979), "Reversibility and Stochastic Networks," John Wiley & Sons, New York 1979.

[7] J. F. C. Kingman (1962), "The Effect of Queue Discipline on Waiting Time Variance," Proc. Camb. Phil. Soc., pp 163-164, 1962.

[8] R. C. Larson (1987), "Perspective on Queues: Social Justice and the Psychology of Queueing," *Oper. Res.* 35, 895-905.

[9] I. Mann (1969), "Queue Culture: The Waiting line as a Social System," Am. J. Sociol, 75, 340-354.

[10] C. Palm (1953), "Methods of Judging the Annoyance Caused by Congestions," TELE (English ed.) No. 2, 1-20.

[11] A. Parekh (1992), "A generalized processor sharing approach to flow control in Integrated services networks," PhD dissertation, MIT, February 1992.

[12] 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*, Vol.1, June 1993, 344-357.

[13] 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*, Vol. 2, 137-150.

[14] J. Rexford, A. Greenberg, F.Bonomi (1996), "Hardware-Efficient Fair Queueing Architectures for High-Speed Networks", IEEE INFOCOM'96, March 1996, 638-646.

[15] M. H. Rothkopf and P. Hech (1987), "Perspectives on Queues: Combining Queues is not Always Beneficial," *Operations Research*, Vol. 35, No. 6, November-December 1987.

[16] Y.T. Wang and R.J.T. Morris (1985), "Load sharing in distributed systems," *IEEE Tx. on computers* Vol. C-34, No. 3, pp. 204-217.

[17] W. Whitt (1984), "The Amount of Overtaking in a Network of Queues," *Networks*, Vol. 14, No. 3, 411-426.

A Glossary of Notation

C_i – Customer number i in order of arrival $i, = 1, 2, ...$

A_i – Arrival time of C_i, with corresponding realization a_i.

S_i – Service time of C_i with corresponding realization s_i.

π – A realization of service order, (a permutation of the indices $1, 2, \ldots$).

δ_i^π – The realized order displacement of C_i in π.

Δ_i – The order displacement (random variable) of C_i for a given discipline.

d_i^π – The waiting displacement of C_i in a realization with order π.

D_i – Waiting displacement (random variable) of C_i for a given discipline, namely $D_i = W_i - W_i^*$ where W_i^* is the waiting time for FIFO.

Φ_1 – The class of disciplines where service selection is done without consideration of the service times of the waiting customers.

Φ_2 – The class of disciplines where the service selection decision can account for the service times of the waiting customers.