Fairness Considerations of Scheduling in Multi-Server and Multi-Queue Systems

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

Benjamin Avi-Itzhak
RUTCOR
Rutgers, the State University of New Jersey
Piscataway, NJ, USA
Email: aviitzha@rutcor.rutgers.edu

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

***Abstract*— Multi-server and multi-queue architectures are common mechanisms used in a large variety of applications (call centers, Web services, computer systems). One of the major motivations behind common queue operation strategies is to grant fair service to the jobs (customers). Such systems have been thoroughly studied by Queueing Theory from their performance (delay distribution) perspective. However, their fairness aspects have hardly been studied and have not been quantified to date. In this work we use the Resource Allocation Queueing Fairness Measure (RAQFM) to quantitatively analyze several multi-server systems and operational mechanisms. The results yield the relative fairness of the mechanisms as a function of the system configuration and parameters. Practitioners can use these results to *quantitatively* account for system fairness and to weigh efficiency aspects versus fairness aspects in designing and controlling their queueing systems. In particular, we quantitatively demonstrate that: 1) Joining the shortest queue increases fairness, 2) A single "combined" queue system is more fair than "separate" (multi) queue system and 3) Jockeying from the head of a queue is more fair than jockeying from its tail.**

I. Introduction

Multiple server queueing models have been used in a large variety of applications, including computer systems, call centers and Web servers, as well as human waiting lines (e.g., in airports). The configuration and operation of multi-server (multi-processor) systems involves a variety of operational mechanisms that must be chosen. These include the number of queues and their dedication, the service policy, the queue joining policy and queue jockeying rules.

An important motivation (perhaps the most important one) behind these mechanisms is the wish to provide *fair service* to the jobs. Fairness in queues was stated to be an important issue (for example [1]–[5]) and this importance (in the eyes of customers) was reinforced by psychological experimental queueing studies [6], [7]. These experimental studies have further shown that the single-queue system is perceived to be more just (fair) compared to the multi-queue system.

Despite the importance of fairness, Queueing Theory, which devoted much effort to the analysis of the above mentioned mechanisms, has dealt mainly with their performance (in terms of delay and its moments) and the analysis of their relative fairness was not carried out. The interest in computer job scheduling and in their fairness has recently raised interest in *quantitatively evaluating queue scheduling fairness*. See [8] for an overview of works done in this area.

The purpose of this work is to use an analytic model, and evaluate the relative fairness of multiple server systems. Such analysis will provide measures of fairness for these systems, that can be used to *quantitatively* account for fairness when considering alternative designs. The quantitative approach can enhance existing design approaches in which usually efficiency (e.g., utilization and delays) is accounted for quantitatively, while fairness is accounted for only in a qualitative way. Particular design issues that can be addressed by this analysis include: 1) Whether to combine queues or not, 2) Whether to allow queue jockeying, and 3) Queue joining policy.

The importance of this paper is in several dimensions. First, it is one of the very first to provide quantitative analysis of fairness for multi-queue multi-server systems; Most other fairness studies were limited to single server systems ([9] is the only exception we are aware of). Second, it approves, in a quantitative manner, several rules of thumb that have been widely believed by practitioners and theoreticians but for which no analytic quantitative support was given. Third, it provides practitioners with a useful quantitative tool by which they can evaluate the fairness of a variety of very common operational strategies.

To carry out this analysis we use the *Resource Allocation Queueing Fairness Measure (RAQFM)* introduced in [10] for single server systems. This measure is based on the basic principle that all customers present at the system at epoch t deserve equal service at that epoch, and deviations from that principle result in discrimination (positive or negative); a summary statistics of these discriminations yields a measure of unfairness. An important property of RAQFM is that it accounts for the tradeoff between *job seniority* and *job size* (see [11]). As such it differs from alternative queue fairness measures, e.g. the ones proposed in [12] and [9], which focus only on seniority or size.

A detailed description of RAQFM, the model, and the notation is given in Section II.

We start our analysis (Section III) by considering G/D/m type systems, that is, systems with a general arrival process and deterministic service times. Under this setting we first

show (Section III-A) that Global FCFS[1], namely serving the customers according to their order of arrival, is the most fair service policy. This holds for any arrival process. This property was proved in [11] for single server systems and is generalized here for multiple server, multiple queue systems.

Second (Section III-B), we study the effect of multiple queues on fairness. This is commonly referred to in the literature as the "Combining Queues" problem, in which a system consisting of m queues, each having service rate μ and arrival rate λ (and denoted the *multi-queue system* or the *separate queue system*) is compared to a system with one queue served by m servers with same service rate μ, and a corresponding arrival rate $m\lambda$ (denoted the *single queue system* or the *combined queue system*). It is widely known that the single queue system is more efficient, i.e. it has a lower mean delay (a proof is given in [13]). However, if jockeying is allowed when a server is idle (in a manner that is nondiscriminatory with respect to the required service time), the systems have the same mean waiting time. In other cases, for example when jockeying favors short jobs, the single queue system might even become less efficient than the separate queue system (see [4] for a full discussion in the matter). As mentioned above, in [6], [7] an experimental psychology study showed that a single queue system is perceived to be more fair than a multiple queue one. Our goal is to quantitatively compare the fairness of the combined queue G/D/m system to that of the separate queue G/D/m system. We show that indeed the combined (single) queue system is more fair than the separate (multiple) queue system.

Third, we investigate the jockeying-on-idle mechanism, namely jockeying is allowed from a non-empty queue to an idle server, and conjecture that allowing for jockeying increases the G/D/m system fairness. We show that jockeying from the head of the queue is more fair than jockeying from its tail (the latter practice is common in some supermarkets).

Fourth, we study (Section III-C) the effect of the queue joining policy on the fairness of the system. Again, the efficiency aspect was thoroughly studied, starting with [14] where the Shortest Queue (SQ) policy is shown to be optimal for exponential arrivals and service times. Further work was done on extending the results for more general systems, for example [15], where two identical exponential servers are considered (with similar results) and [16] where a more general system is considered. Some cases where SQ is not optimal are given in [17]. We study the fairness of the G/D/m system under SQ, Round-Robin (RR), and Random Queue Joining (RAND). We show that SQ and RR are equivalent to each other and they both are more fair than RAND.

Lastly for G/D/m type systems (III-D), we examine simulation results of these systems under Poisson arrivals. The results support all the properties shown.

We then (Section IV) turn to study M/M/m type systems, namely systems with Poisson arrivals and exponential service times. We provide an exact fairness analysis for the same systems discussed in Section III-B and Section III-C, for M/M/m type systems. The analysis of each of these systems leads to a set of linear equations that needs to be solved numerically to evaluate the fairness (unfairness) of the system. Numerical evaluation of these equations (also supported by simulation) for a wide set of load parameters demonstrates that a) the properties proved for the G/D/m type systems in Section III-B hold for M/M/m type systems as well. b) the queue joining policy has little effect on the unfairness for M/M/m type systems.

We then (Section V) turn to examine G/G/m type systems. We show, via a simple counter-case, that the results derived in Section III and Section IV do not necessarily hold for any arbitrary G/G/m model. We provide sufficient conditions under which they hold for the G/GI/m model.

Practical implications are discussed in Section VI, and lastly, concluding remarks are given in Section VII.

II. System Model and RAQFM

Consider a queueing system with m servers, indexed $1, 2, \ldots, m$. All servers have equal service rate, for simplicity a rate of one (unit of service time per unit of time). The system is subject to the arrival of a stream of customers, $C_1, C_2, \ldots$, who arrive at the system at this order. In order to focus on the effect multiple servers and multiple queues have on the system fairness we assume all customers belong to the same class and have the same priority.

Let a_l and d_l denote the arrival and departure epochs of C_l respectively. Let s_l denote the service requirement (measured in time units) of C_l. Let g_l denote the epoch in which C_l starts being served.

Corresponding to the way RAQFM was previously described in [10], [11], [18], the unfairness in the system is evaluated as follows: The basic fundamental assumption is that at each epoch, all customers present in the system deserve an equal share of the *service given*. If we let $0 \leq \omega(t) \leq m$ denote the total service rate given at epoch t (which usually is an integer equaling the number of working servers at that epoch), and $N(t)$ denote the number of customers in the system at epoch t, then the fair share, called the *momentary warranted service rate*, is $\omega(t)/N(t)$. Let $\sigma_l(t)$ be the momentary actual rate at which service is given to C_l at epoch t. This is called the *momentary granted service rate*. The *momentary discrimination rate* of C_l at epoch t, denoted $\delta_l(t)$, is therefore $\delta_l(t) = \sigma_l(t) - \omega(t)/N(t)$. This can be viewed as the rate at which customer discrimination accumulates for C_l at epoch t. This is a generalization (proposed in [18]) of the definition for the single server case, given in [10], [11].

The total discrimination of C_l, denoted D_l, is

$$D_l = \int_{a_l}^{d_l} \delta_l(t)dt. \tag{1}$$

Let D be a random variable denoting the discrimination of an arbitrary customer when the system is in equilibrium. As shown in [18], the expected value of discrimination always

[1]In this entire paper, when we say FCFS we refer to Global FCFS, and similarly for LCFS

obeys $E[D] = 0$. Thus, according to RAQFM, the unfairness of the system is defined as the second moment of the discrimination, namely $E[D^2]$, and denoted F_{D^2}.

Throughout this paper we use ρ in its common definition as the system utilization factor, which is $\rho \stackrel{def}{=} \lambda\bar{x}/m$ where λ is the average arrival rate and $\bar{x}$ is the average service time (e.g. [19, Sec. 2.1]).

Remark 2.1 (Alternative Measure): Alternatively to setting $\omega(t)$ to be the amount of service actually given at epoch t, one can set it to be the number of servers $\omega(t) = m$ or the number of servers that can be utilized (depending on the number of customers present) $\omega(t) = \min(N(t), m)$. This will interpret an idle server as a source of negative discrimination (which is not compensated by concurrent positive discrimination). We do not investigate this measure here and it is the subject of future research.

Remark 2.2 (Ultimately Fair Systems): Note that for the Processor Sharing policy and the G/G/∞ system all customers have zero discrimination and thus the system is ultimately fair. The first since the momentary service rate given to each customer is $\sigma_l(t) = \omega(t)/N(t)$ and thus the momentary discrimination is $\delta_l(t) = \omega(t)/N(t) - \omega(t)/N(t) = 0$, and the second since $\omega(t) = N(t)$ and $\sigma_l(t) = 1$, thus $\delta_l(t) = 1 - N(t)/N(t) = 0$.

III. Fairness of G/D/m Type Systems

In this section we analyze the fairness of G/D/m type systems, i.e. the customers arrive according to a general arrival process, the service requirements of all customers are identical, and the system has m servers.

A. The Effect of Seniority on Fairness

We start with analyzing the effect of customer seniority on fairness, and show that under the G/D/m model serving customers in order of seniority increases system fairness.

For simplicity we choose to demonstrate this for adjacently served customers. Consider two customers C_j and C_k that are adjacently served, i.e. they are scheduled to begin service adjacently, with arrival times $a_j < a_k$ and equal service requirements, i.e. $s_j = s_k$. Observe two possible cases: In the first case (Fig. 1) they are served sequentially, by the same server. In the second case (Fig. 2) they are served by two different servers in a partially-parallel manner.

We have two possible schedules: In schedule (a), (Fig. 1(a), Fig. 2(a)), the order of seniority is preserved, i.e. C_j is served before C_k. In schedule (b), (Fig. 1(b), Fig. 2(b)), the order of service of C_j and C_k is interchanged and thus the order of seniority is violated. We assume that both of the schedules are possible, i.e. that C_j and C_k reside in the queue together for some time.

Lemma 3.1 (Preference of Seniority): Let C_j and C_k be two customers, where $a_j < a_k$ and their service times are equal, i.e. $s_j = s_k = s$. Then the unfairness (measured as F_{D^2}) of the seniority preserving schedule (a) is smaller than the unfairness of the seniority violating schedule (b), for every arrival pattern.

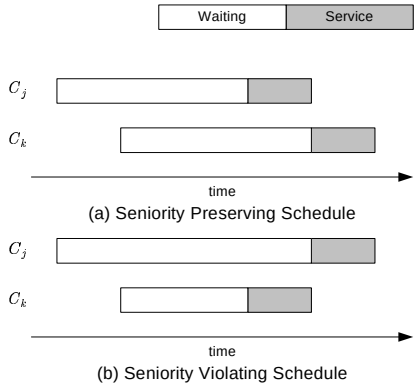

Fig. 1. Case 1 - Sequentially Served Customers (By The Same Server)

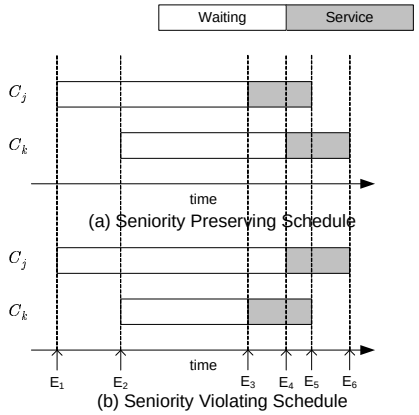

Fig. 2. Case 2 - Partially Parallel Served Customers

Proof: For simplicity we first prove this for adjacently served customers.

We need to consider both cases. We note that the first case (Fig. 1) is the only case which was considered for single server systems in [11, Theorem 3.1], and it was proven there that this lemma holds. We therefore only need to prove the second case (Fig. 2).

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

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

Define ΔF_{D^2}, $\Delta\hat{F}$, and $\Delta\tilde{F}$ to be the change, due to the interchange between C_j and C_k, in the values of F_{D^2}, $\hat{F}$, and $\tilde{F}$ respectively. We need to prove that $\Delta F_{D^2} \geq 0$, where

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

where L is the number of customers in the observed scenario.

We divide the time interval $(a_j, \max(d_j, d_k))$ (namely from the arrival of C_j until both C_j and C_k depart) into five sub-intervals, where interval $i, i = 1, 2, 3, 4, 5$, is (E_i, E_{i+1}), and where $E_1 = a_j$; $E_2 = a_k$; E_3 is the first point in time where service to either C_j or C_k starts; E_4 is the point in time where service to the second customer starts; $E_5 = E_3 + s$; $E_6 = E_4 + s$ (see Fig. 2). As customers are adjacently served, $E_5 > E_4$.

We note that the number of customers in the system under schedule (a), $N^a(t)$, and the number of customers in the system under schedule (b), $N^b(t)$, are identical for every t. Therefore, the warranted service in interval i (of any customer) is the same in both schedules, and we denote it $R_{(i)}$. As the warranted service in each interval is the same in both schedules, and as other customers' departure epochs are also the same in both schedules, the interchange of C_j and C_k affects only C_j and C_k, i.e. $D_i^a = D_i^b, i \neq j, k$. Therefore $\Delta\tilde{F} = 0$, and from (2) it follows that

$$\begin{aligned}\Delta F_{D^2} = \Delta\hat{F} &= \frac{1}{L}\left[(D_j^b)^2 + (D_k^b)^2 - (D_j^a)^2 - (D_k^a)^2\right] \\ &= \frac{1}{L}\left[\left(s - (R_{(1)} + R_{(2)} + R_{(3)} + R_{(4)} + R_{(5)})\right)^2\right. \\ &\quad + \left(s - (R_{(2)} + R_{(3)} + R_{(4)})\right)^2 \\ &\quad - \left(s - (R_{(1)} + R_{(2)} + R_{(3)} + R_{(4)})\right)^2 \\ &\quad \left. - \left(s - (R_{(2)} + R_{(3)} + R_{(4)} + R_{(5)})\right)^2\right] = \frac{2}{L} R_{(1)} R_{(5)} > 0,\end{aligned} \tag{3}$$

where the inequality is due to $R_{(i)} > 0$, $i = 1, 2, 3, 4, 5$.

This completes the proof for adjacently served customers. To address non adjacently served customers one has to address a third case, similar to case (a) where the service of the second customer starts after the service completion of the first customer. The correctness of this case is very similar to that of case (a). ■

Remark 3.1: Lemma 3.1 holds for $s_j = s_k = s$ and regardless of the service times of the other customers.

Define Φ to be the class of non-preemptive, non-divisible service policies (i.e. service policies where once the server started serving a customer it will not stop doing so until the customer's service requirement is fulfilled, and at most one customer is served at any epoch by each server), where the scheduler does not know the actual values of the service times, or does not account for them in the scheduling decisions.

Theorem 3.1 (Fairness of FCFS under G/D/m): If the service requirements of all customers are identical, then for every arrival pattern, FCFS is the least unfair service policy in Φ and LCFS is the most unfair one.

Proof: The proof is similar to the proof used in [11, Theorem 3.3] for single server systems.

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

Given this arrival pattern and the order of service created by ϕ, identify the first pair of customers which are adjacently served and are not served according to their order of arrival. Interchange the order of service between these two customers (which is certainly possible). According to Lemma 3.1, the result of this interchange is a decrease in the overall unfairness. Thus the resulting order of service is more fair than ϕ, in contradiction to ϕ being the least unfair service policy for this arrival pattern.

A similar argument proves that LCFS is the most unfair policy in Φ. ■

B. The Effect of Multiple Queues Operation Mechanisms on Fairness

In this section we study several design and service decision considerations and examine their effect on the system fairness.

The first practical issue is the use of multiple queues versus the use of a single queue (also called sometimes 'separate' and 'combined' queues). In many cases both options are physically possible, and the selection between them is done mainly based on the supposed efficiency of the system (see [4]).

The second issue, assuming multiple queues are used, is jockeying. In some systems jockeying is possible and we would like to find the effect it has on the system fairness. We focus on jockeying done while a server is idle, in a manner that is nondiscriminatory with respect to the required service time.

The third issue, assuming jockeying is possible, is whether jockeying should be done from the head of the queue or from its tail. Note that from the efficiency (mean delay) point of view there is no difference between the two manners, as long as the jockeying is nondiscriminatory with respect to the required service time.

1) Multiple Queues Systems' Description: For the sake of illustration, we analyze four systems. All four systems have m servers. The first system has one common queue. A customer arriving at the system joins the tail of the queue, and when a server becomes free it serves the customer at the head of the queue, thus the system is non-idling[2]. The customers are therefore admitted into service according to their order of arrival in a global FCFS manner. We refer to this system as the *single queue system*.

The second system has m queues, where each of the servers serves only customers joining its assigned queue, in a FCFS manner. Customers arriving at the system choose one of the queues, or are assigned to one of the queues, and jockeying between the queues is not permitted at any epoch. We refer to this system as the *standard multiple queue system*. In this section we address a system where queue assignment (or joining) is done randomly. In Section III-C we address the effect the queue joining policy has on the fairness.

The third and fourth systems have m dedicated FCFS queues, and jockeying is allowed at epochs when a server

[2]A system is *non-idling* if the number of busy servers (and thus the actual service rate) is always $\min(N, m)$, where N is the number of customers in the system.

is idle. When this happens, a customer from one of the non-empty queues is immediately assigned to that server. Thus, as long as there are at least m customers in the system, none of the servers is ever idle. We consider two manners in which the jockeying customer is chosen, either from the head of one of the non-empty queues or from the tail of one. We refer to the first as the *head-jockeying-on-idle* multi-server system and to the second as the *tail-jockeying-on-idle* system. In both systems, if there are several non-empty queues, then the queue from which the jockeying is done is chosen randomly. For simplicity we do not address other manners in which that queue can be decided, but note that this manner is the most probable in systems where the customers choose to jockey on their own.

As mentioned in Section I, both jockeying-on-idle systems are known to be as efficient as the single queue system, while the standard system is known to be less efficient than the other three systems.

2) Multiple Queues Systems' Properties:

Theorem 3.2 (Single Queue is Fair): If the service requirements of all customers are equal, then for every arrival pattern, the single queue system has the lowest unfairness among the systems mentioned above.

Proof: The proof is immediate since serving in a single queue means serving in a global FCFS manner, which is proven in Theorem 3.1 to be the most fair policy. ■

Theorem 3.3 (Head-Jockeying-On-Idle is Fair): If service requirements of all customers are equal, and the choice of queue from which jockeying is done is identical, then for every arrival pattern the head-jockeying-on-idle system has lower unfairness than the tail-jockeying-on-idle system.

Proof: We prove that if the choice of queue from which jockeying is done is identical, the policy with the lowest unfairness is one where jockeying is always done from the head of the queue. Specifically, this proves that head-jockeying is more fair than tail-jockeying.

We prove this by way of contradiction. Assume for the contradiction that there exists a jockeying policy ψ by which the jockeying customer is not always the customer at the queue head, and that ψ is the policy with the lowest unfairness for every arrival pattern. We will show that a policy with lower unfairness can be constructed.

Consider an arrival pattern where there exists an epoch e in which a server was idle, and in which the customer admitted to the idle server by ψ was the n-th customer in some non-empty queue, and $n > 1$. Say there are N customers in that queue, and denote the customers in that queue $C_1, C_2, \ldots, C_N$, according to their order of arrival, with C_1 being the first to arrive. We show that admitting $C_n, n > 1$ to the idle server has higher unfairness than admitting C_{n-1}.

Say C_n was admitted to the idle server at epoch e, and C_{n-1} was served at some later epoch $e' > e$. We construct a policy ψ' that exactly imitates ψ, except it exchanges the roles of C_n and C_{n-1}, i.e. C_{n-1} is admitted at epoch e and C_n is served at epoch e'. Since the customers are adjacent in the queue and have equal service requirements no other customers are influenced by this exchange. Therefore, using the same notations and arguments used in Lemma 3.1, $\Delta F = \Delta\hat{F}$ and we need only to prove that $\Delta\hat{F} > 0$. For brevity and comparability with the proof of Lemma 3.1 we denote $j = n-1, k = n$.

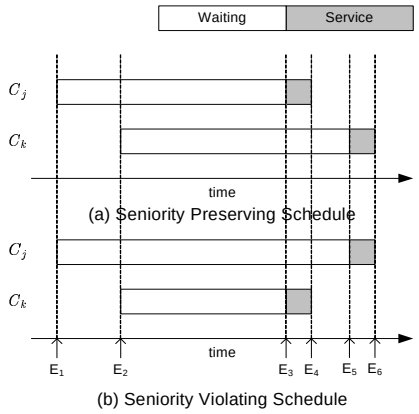

Fig. 3. Selection of Jockeying Customer, Equal Service Requirements, Case 2.

Let the service requirement of C_j and C_k be s. Two cases should be considered: either $e' - e < s$ or $e' - e \geq s$. For $e' - e < s$ observe that the situation is exactly the one depicted in Fig. 2, and as shown in the proof of Lemma 3.1 $\Delta\hat{F} > 0$ in this case.

For the second case, $e' - e > \tau$, the situation is depicted in Fig. 3. We divide the time interval $(a_j, e' + s)$ (namely from the arrival C_j until both C_j and C_k depart) into five sub-intervals where interval $i, i = 1, 2, 3, 4, 5$, is (E_i, E_{i+1}), where $E_1 = a_j$; $E_2 = a_k$; E_3 is the jockeying epoch e; $E_4 = E_3 + s$; E_5 is the epoch where the second customer is served, e'; $E_6 = E_5 + s$ (see Fig. 3).

Let the service warranted in interval i be $R_{(i)}$. Using (2) the difference in unfairness is

$$\begin{aligned}\Delta\hat{F} = \frac{1}{L}\big[&\big(s - (R_{(1)} + R_{(2)} + R_{(3)} + R_{(4)} + R_{(5)})\big)^2 \\ &+ \big(s - (R_{(2)} + R_{(3)})\big)^2 \\ &- \big(s - (R_{(1)} + R_{(2)} + R_{(3)})\big)^2 \\ &- \big(s - (R_{(2)} + R_{(3)} + R_{(4)} + R_{(5)})\big)^2\big] \\ &= \frac{2}{L}R_{(1)}(R_{(4)} + R_{(5)}) > 0. \quad (4)\end{aligned}$$

We therefore showed that for every service policy ψ that performs jockeying not always from the head of the queue for at least one arrival pattern, a better policy ψ' exists, in contradiction to ψ being the policy with the lowest unfairness for every arrival pattern. Therefore, the policy with the lowest unfairness for every arrival pattern must be one where the jockeying customer is always the one at the queue head. ■

Remark 3.2: We have also shown that the worst policy is a policy in which the customer admitted is always the one at the queue tail.

Conjecture 3.1 (jockeying-On-Idle is Fair): If the service requirements of all customers are equal, then for every arrival pattern the standard system is less fair than the jockeying-on-idle systems.

We base this conjecture on multiple simulations and on intuition. Intuitively speaking, allowing jockeying allows more servers to work simultaneously, thus distributing the service rate between more customers.

Remark 3.3 (Non-Equal Service Rate): As mentioned in Section II we limit our model to servers with the same service rate (for simplicity a rate of one). It can be easily shown that when this is not the case, the properties discussed above are not necessarily true. For example, a more fair policy than FCFS might hold a senior customer in order to serve it by a faster server.

C. The Effect of Queue Joining Policy on Fairness

In this section we address the effect of the queue joining on fairness. To this end we analyze the head-jockeying-on-idle system with three queue joining policies: (a) customers join queues at random (RAND) (b) customers always join the shortest queue (SQ), and (c) customers join queues in a round-robin manner (RR).

Note that in practice selection of the queue joining policy can be affected by several physical factors of the system. These include who is making the queue joining decision (the customer or a queue manager) and whether queue size information is available to the decision maker.

Theorem 3.4 (Queue Joining Policies): If all service requirements are equal, then the unfairness in SQ equals that in RR, and both have the lowest unfairness.

Proof: To prove that the unfairness for SQ equals that of RR note that if service requirements are equal, SQ and RR are in fact identical, since using round-robin always directs a customer to the shortest queue.

Second, note that using SQ (or RR) yields an order of service which is in fact identical to global FCFS, and from Theorem 3.1 Global FCFS is the most fair order of service. ■

D. Numerical Results for the M/D/2 Model

Fig. 4 depicts numerical results from simulating the M/D/2 model as a function of the system utilization factor $\rho = s\lambda/2$ for the four systems discussed in Section III-B.1, where the queue to join is chosen randomly. Service requirement was set as one unit ($s = 1$). Each point is the result of simulating the passage of at least 10^6 customers through the system. The figure demonstrates the following properties:

1. Except for one system and one point (single queue system at high load) the unfairness is monotone increasing with the system utilization factor ρ. This monotonic behavior was demonstrated several times before (see [10], [11], [18]). The reason for the decrease in unfairness at high loads for the single queue system is still unclear and further study is required.

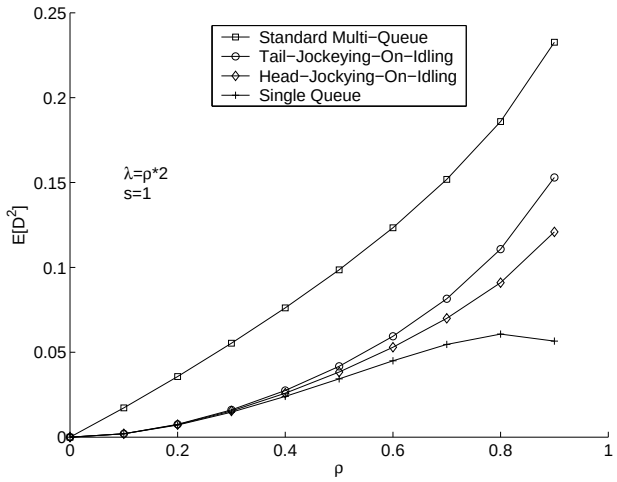

Fig. 4. Unfairness of Four Queue Strategies under the M/D/2 Model

2. The single queue system is more fair than all the multiple queue systems, for every system load, as expected from Theorem 3.2. The ratio between the unfairness of the single queue system and that of the head-jockeying-on-idle system increases with ρ, and reaches a ratio of more than $1:2$ for $\rho = 0.9$. Recall that as long as jockeying is allowed only in a manner that is nondiscriminatory with respect to the required service time (as is the case with the multiple queue systems analyzed) the single queue system is also at least as efficient as the multiple server systems.

3. Among the multiple queue systems, the jockeying-on-idle systems are more fair than the standard (idling) system, for every system load. The ratio between the unfairness of the standard system and that of the head-jockeying-on-idle system decreases with ρ, starting with a high ratio of more than $1:8$ for $\rho = 0.1$ and reaching a low of $1:1.9$ for $\rho = 0.9$. This agrees with Conjecture 3.1.

4. Among the jockeying-on-idle systems, head-jockeying is more fair than tail-jockeying for every system load, as expected from Theorem 3.3. The difference between them increases with ρ and is as high as 25% for $\rho = 0.9$. Recall that these two systems are equally efficient.

In Fig. 5 the three queue joining policies discussed in Section III-C (head-jockeying-on-idle) are compared to each other, as a function of the system utilization factor ρ, for the M/D/2 model. Service requirement was set as one unit. Each point is the result of simulating the passage of at least 10^6 customers through the system. The figure demonstrates the following properties:

1. SQ and RR are identical, as expected from Theorem 3.4. In fact, they are identical to the single queue M/D/2 system.

2. RAND has higher unfairness than SQ and RR for every system load, as expected from Theorem 3.4. The difference between them increases with ρ and is as high as 90% for $\rho = 0.9$. Recall that these two systems are equally efficient.

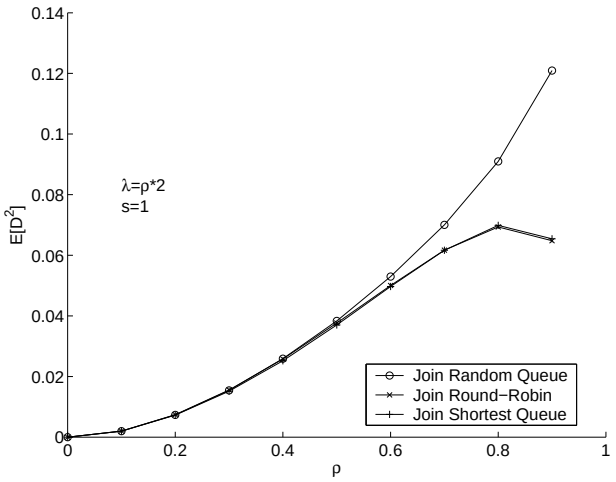

Fig. 5. Unfairness in the M/D/2 Model - Comparison of Queue Joining Policies

IV. Fairness of M/M/m Type Systems

In this section we analyze the fairness of M/M/m type systems. We start with a description of the assumptions and notations used in our analysis.

For the analysis in this section we will consider systems where the arrival of customers follows a Poisson process with rate λ and the service requirements of the customers are i.i.d. exponentially with mean $1/\mu$. Let $\rho \stackrel{def}{=} \lambda/(m\mu)$ be the fraction of busy servers (also called the system utilization factor). For stability we assume $\rho < 1$.

Due to the Markovian nature of the system, and for mathematical convenience, time can be viewed as being slotted where the sequence of arrivals and departures forms the slot boundaries. Let $T_i, i = 0, 1, \ldots$ be the duration of the i-th slot.

We limit the analysis to systems where a service decision is made only on arrival and departure epochs. Thus, the number of available servers, the number of servers actually giving service, and the rate at which service is given to each customer, are constant during each slot. We can therefore define $0 \leq \omega_i \leq m$ as the number of working servers in the i-th slot, $\sigma_{i,l}$ as the rate at which service is given to C_l at the i-th slot, and N_i as the number of customers in the system during the i-th slot. $\delta_{i,l}$, the momentary discrimination of C_l during the i-th slot, which is the rate at which customer discrimination accumulates for C_l at this slot, is $\delta_{i,j} = \sigma_{i,l} - \omega_i/N_i$. The total discrimination accumulated for C_l during the i-th slot is $c_{i,l}T_i$. Thus, the slotted version of (1) is $D_l = \sum_{i=x_l}^{y_l} \delta_{i,l}T_i$, where x_l is the index of the slot initiated by the arrival of C_l and y_l is the index of the slot terminated by the departure of d_l.

A. The Effect of Multiple Queues on Fairness

In this section we provide quantitive fairness analysis, under the M/M/m model, for the four systems described in Section III-B.1, namely the single queue system, the standard multiple queue system, and the two jockeying-on-idle systems - head-jockeying-on-idle and tail-jockeying-on-idle. Recall that for all four systems the order of service whithin each queue is FCFS, except for jockeying.

1) Fairness in Multiple Server Single Queue Systems under the M/M/m Model: Consider a single queue system, as described in Section III-B.1. Consider an arbitrary tagged customer denoted C. Let a be the number of customers ahead of C in the queue, including served customers. If C is in service, $a \stackrel{def}{=} 0$. Let b be the number of customers behind C. If C is in service, b includes also customers served by other servers. Note the (unavoidable) jump occurring in the value of b when C enters service. Due to the Markovian nature of the system, and due to non-idling, the state (a, b) captures all that is needed to predict the future of C.

The momentary discrimination during a slot where C is in state (a, b), denoted $\delta(a, b)$, is $\delta(a, b) = \mathbb{1}(a = 0) - s/(a+b+1)$, where s is the number of customers served at that state, $s = min(m, a + b + 1)$, and $\mathbb{1}(\cdot)$ is the logical function (also called sometimes an indicator function), i.e. $\mathbb{1}(\cdot) = 1$ if the relation $\cdot$ is true and $\mathbb{1}(\cdot) = 0$ otherwise.

We would like to compute $E[D^2]$, the unfairness of the system, as define in Section II. Let $E[D^2|k]$ for $k = 0, 1, \ldots$ denote the expected value of the square of discrimination, given that the customer encounters k customers on arrival (including the ones being served). Let p_k be the steady state probability that there are k customers in the system. According to the PASTA property (see [20]), this is also the probability that k customers are seen by an arbitrary arrival. Thus, the second moment of D (the unfairness) follows

$$E[D^2] = \sum_{k=0}^{\infty} E[D^2|k]p_k. \tag{5}$$

The steady state probabilities for the single queue M/M/m system are:

$$\begin{aligned} p_k &= \begin{cases} p_0 \frac{(m\rho)^k}{k!} & k \leq m \\ p_0 \frac{\rho^k m^m}{m!} & k \geq m \end{cases} \\ p_0 &= \left[\frac{(m\rho)^m}{m!(1-\rho)} + \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} \right]^{-1}, \end{aligned} \tag{6}$$

(e.g. [19, Sec. 3.5]).

In the case $m = 2$ these are equal to $p_k = p_0 2\rho^k \quad k \geq 1$, $p_0 = (1 + \rho)/(1 - \rho)$.

Let $D(a, b)$ be a random variable denoting the discrimination experienced by a customer, through a walk starting at (a, b), and ending at its departure. Then

$$E[D^2|k] = \begin{cases} E[D^2(k, 0)] & k \geq m \\ E[D^2(0, k)] & k < m \end{cases}. \tag{7}$$

Let $d(a, b)$ and $d^{(2)}(a, b)$ be the first two moments of $D(a, b)$.

In the single queue $M/M/m$ system the slot lengths are exponentially distributed with parameter $\lambda + s\mu$ and first two moments $t^{(1)}(a, b) = 1/(\lambda + s\mu)$, $t^{(2)}(a, b) = 2(t^{(1)})^2$.

Define $\tilde{\lambda}(a, b)$, the probability that the slot will end with a customer arrival event, and $\tilde{\mu}(a, b)$, the probability that the

slot will end with a customer departure from a specific active server. Then $\tilde{\lambda}(a,b) = \lambda/(\lambda+s\mu)$, $\tilde{\mu}(a,b) = \mu/(\lambda+s\mu)$. Note that here and throughout the paper $\tilde{\lambda}$ refers to the probability of an arrival of *any* customer, while $\tilde{\mu}$ refers to the probability of a departure of a customer from *one specific* queue. This seeming inconsistency is required for mathematical brevity.

Assume C is in state (a,b). At the slot's end, the system will encounter one of the following events and C's state will change accordingly:

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

2. For $a>0$: A customer leaves the system. The probability of this event is $s\tilde{\mu}(a,b)$. If $a>m$, C's state changes to $(a-1,b)$. Otherwise, C's state changes to $(0,a+b-1)$.

3. For $a=0, b>0$: A customer other than C leaves the system. The probability of this event is $(s-1)\tilde{\mu}(a,b)$. C's state changes to $(a,b-1)$.

4. For $a=0$: C leaves the system. The probability of this event is $\tilde{\mu}(a,b)$.

This leads to the following set of linear equations

$$d(a,b) = t^{(1)}(a,b)\delta(a,b) + \tilde{\lambda}(a,b)d(a,b+1) + \tilde{\mu}(a,b)\begin{cases} md(a-1,b) & a>m \\ md(0,a+b-1) & a=m \\ (s-1)d(a,b-1) & a=0,b>0 \end{cases} \quad (8)$$

$$d^{(2)}(a,b) = t^{(2)}(a,b)(\delta(a,b))^2 + \tilde{\lambda}(a,b)d^{(2)}(a,b+1) + \tilde{\mu}(a,b)\begin{cases} md^{(2)}(a-1,b) & a>m \\ md^{(2)}(0,a+b-1) & a=m \\ (s-1)d^{(2)}(a,b-1) & a=0,b>0 \end{cases} + 2t^{(1)}(a,b)\delta(a,b)\left(\tilde{\lambda}(a,b)d(a,b+1) + \tilde{\mu}(a,b)\begin{cases} md(a-1,b) & a>m \\ md(0,a+b-1) & a=m \\ (s-1)d(a,b-1) & a=0,b>0 \end{cases}\right). \quad (9)$$

These sets of equations can be solved numerically to any required precision using simple numeric methods. Solving these sets of equations yields $d^{(2)}(a,b)$, and substituting the results in (7) yields $E[D^2|k]$. Substituting these in (5) yields $E[D^2]$, the system unfairness.

2) Fairness in Multiple Server Multiple Queue Systems: We now analyze the standard multiple queue system, as described in Section III-B.1. Consider an arbitrary tagged customer C. When C arrives into the system it joins a queue at random. We refer to the queue that the customer joined as the *local* queue and to the rest of the queues as the *other* queues. For the analysis we index the local queue 1, and the other queues $2,\ldots,m$, in some arbitrary way.

Let a be the number of customers ahead of C in the local queue, including the served customer, if such a customer exists. Let b be the number of customers behind C in the local queue. Let l_i be the number of customers in the i-th queue, $i=2,\ldots,m$, including any served customers, and let $\mathbf{l}$ denote the vector $(l_2,\ldots,l_m)$. Due to the Markovian nature of the system, the state $(a,b,\mathbf{l})$ captures all that is needed to predict the future of C.

Define $\Sigma_{\mathbf{v}}$, the sum of the elements of a vector $\mathbf{v}$. Recall that s denotes the number of active servers (the number of customers served). Since the server at the local queue is active as long as C is in the system

$$s = 1 + \sum_{2\le i\le m} \mathbb{1}(l_i>0). \quad (10)$$

The momentary discrimination during a slot in which C is in state $(a,b,\mathbf{l})$, denoted $\delta(a,b,\mathbf{l})$, is

$$\delta(a,b,\mathbf{l}) = \mathbb{1}(a=0) - \frac{s}{\Sigma_{\mathbf{l}}+a+b+1}. \quad (11)$$

Let $\mathbf{k} = k_1,\ldots,k_m$ denote a queue occupancy, where k_i is the number of customers at the i-th queue, including the customer in service. Let $E[D^2|\mathbf{k}]$ denote the expected value of the square of discrimination, given that the customer encounters occupancy $\mathbf{k}$ upon arrival.

Let $P_{\mathbf{k}}$ be the steady state probability that the occupancy is $\mathbf{k}$, Then

$$E[D^2] = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\cdots\sum_{k_m=0}^{\infty} E[D^2|\mathbf{k}]P_{\mathbf{k}}. \quad (12)$$

Note that the system consists of m independent M/M/1 queues, where each is utilized a fraction $\rho=\lambda/(\mu m)$ of the time and therefore $P_{\mathbf{k}} = (1-\rho)^m\rho^{\Sigma_{\mathbf{k}}}$.

Let $D(a,b,\mathbf{l})$ be a random variable, denoting the discrimination experienced by a customer, through a walk starting at $(a,b,\mathbf{l})$, and ending at its departure. Let $\hat{\mathbf{k}}_i$ denote the vector $\mathbf{k}$, whose i-th element is ommited, i.e. $\hat{\mathbf{k}}_i = (k_1,\ldots,k_{i-1},k_{i+1},\ldots,k_m)$. Using this notation

$$E[D^2|\mathbf{k}] = \frac{1}{m}\sum_{i=1}^{m} E[D^2(k_i,0,\hat{\mathbf{k}}_i)]. \quad (13)$$

Let $d(a,b,\mathbf{l})$ and $d^{(2)}(a,b,\mathbf{l})$ be the first two moments of $D(a,b,\mathbf{l})$.

When C is in state $(a,b,\mathbf{l})$ the slot length is exponentially distributed with parameter $\lambda+s\mu$ and first two moments

$$t^{(1)}(a,b,\mathbf{l}) = \frac{1}{\lambda+s\mu}, \qquad t^{(2)}(a,b,\mathbf{l}) = 2(t^{(1)})^2. \quad (14)$$

Also define $\tilde{\lambda}(a,b,\mathbf{l})$, the probability that the slot will end with a customer arrival and $\tilde{\mu}(a,b,\mathbf{l})$, the probability that the slot will end with a customer departure event from a specific active server. These are equal to

$$\tilde{\lambda}(a,b,\mathbf{l}) = \frac{\lambda}{\lambda+s\mu}, \qquad \tilde{\mu}(a,b,\mathbf{l}) = \frac{\mu}{\lambda+s\mu}. \quad (15)$$

Define $\mathbf{I}_i$, the vector $(l_2=0, l_3=0,\ldots,l_i=1,\ldots,l_m=0)$.

At the slot end, the system will encounter one of the following events and C's state will change accordingly:

1. A customer arrives at the system, joining the i-th queue. The probability of this event is $\tilde{\lambda}(a,b,\mathbf{l})/m$, for every value of $i=1,\ldots,m$, and the sum of these probabilities is therefore $\tilde{\lambda}(a,b,\mathbf{l})$. If $i=1$, C's state changes to $(a,b+1,\mathbf{l})$. Otherwise, C's state changes to $(a,b,\mathbf{l}+\mathbf{I}_i)$.

2. A customer leaves the system, from the i-th queue. This is possible only for $i=2,\ldots,m$ such that $l_i>0$, and for $i=1$ (the local queue). The probability of this event is $\tilde{\mu}(a,b,\mathbf{l})$, for each of the s non-empty queues, and the overall probability of departure is therefore $s\tilde{\mu}(a,b,\mathbf{l})$. If $i=1$ and $a>0$, C's state changes to $(a-1,b,\mathbf{l})$. If $i=1$ and $a=0$ (C is being served), C leaves the system. If $i>1$, C's state changes to $(a,b,\mathbf{l}-\mathbf{I}_i)$.

This leads to the following set of linear equations

$$d(a,b,\mathbf{l}) = t^{(1)}(a,b,\mathbf{l})\delta(a,b,\mathbf{l}) + \tilde{\lambda}(a,b,\mathbf{l}) \frac{1}{m}\big(d(a,b+1,\mathbf{l}) + \sum_{i=2}^{m} d(a,b,\mathbf{l}+\mathbf{I}_i)\big) + \tilde{\mu}(a,b,\mathbf{l}) \Big(\sum_{2\le i\le m} \mathbb{1}(l_i>0)d(a,b,\mathbf{l}-\mathbf{I}_i) + \mathbb{1}(a>0)d(a-1,b,\mathbf{l})\Big). \tag{16}$$

For brevity, we omit the set of linear equations for $d^{(2)}(a,b,\mathbf{l})$. This can be found in [21]. Similarly to Section IV-A.1, solving those equation yields $d^{(2)}(a,b,\mathbf{l})$, substituting in (13) yields $E[D^2|\mathbf{k}]$, and substituting in (12) yields $E[D^2]$, the system unfairness.

3) Fairness in Multiple Server Multiple Queue Systems With Jockeying-On-Idle: We now turn to the two jockeying-on-idle systems described in Section III-B.1. As we already provided full analysis of two systems, we omit some of the details. The full analysis can be found in [21].

First let us define the jockeying-on-idle system behavior in more detail. When a customer arrives to the system finding no idle servers, the customer is assigned to a queue randomly (we discuss other options in Section IV-B). If at least one of the servers is idle, the customer is assigned at random to one of the idle servers. In each queue the order of service is FCFS. In the event that a server becomes idle a customer is assigned to that server from one of the non-empty queues, where the queue is chosen randomly. In the head-jockeying-on-idle system the assigned customer is the customer at the head of the chosen queue and in the tail-jockeying-on-idle system the assigned customer is the customer at its tail.

The analysis of this system is almost identical to that of the standard system analyzed in Section IV-A.2. The state definition is identical. s, the number of customers served, can be calculated in a simpler way since $s=\min(\Sigma_{\mathbf{l}}+1,m)$, although (10) holds as well. The momentary discrimination, the slot length moments, and the arrival and departure probabilities follow (11), (14), and (15), respectively.

Note that due to non-idling some occupancy vectors are impossible, so $P_{\mathbf{k}}=0$ for $k_i=0,k_j>1,i,j=1,\ldots,m$. For completeness in this case we also define $E[D^2|\mathbf{k}]\stackrel{def}{=}0$.

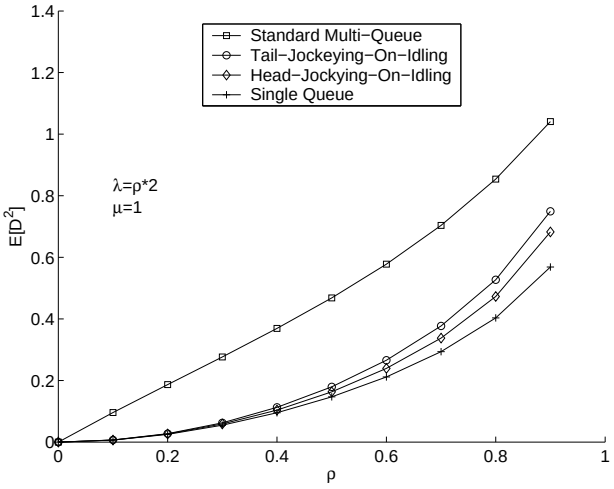

Fig. 6. Unfairness of Four Queue Strategies for the M/M/2 Model

Using these definitions (12) holds, and the steady state probabilities can be numerically calculated using the system's balance equations, which we omit for brevity.

The relationship between $E[D^2|\mathbf{k}]$ and $D(a,b,\mathbf{l})$ is different in the jockeying-on-idle systems, since the arriving customer is assigned to an idle server if such a server exists. If $\Sigma_{\mathbf{k}}<m$ there are $z=m-\Sigma_{\mathbf{k}}$ idle servers and we let e_i be the index of the i-th idle server, $i=1,\ldots,z$. Then

$$E[D^2|\mathbf{k}] = \begin{cases} \frac{1}{m}\sum_{i=1}^{m} E[D^2(k_i,0,\hat{\mathbf{k}}_i)] & \Sigma_{\mathbf{k}}\ge m \\ \frac{1}{z}\sum_{i=1}^{z} E[D^2(0,0,\hat{\mathbf{k}}_{l_i})] & \Sigma_{\mathbf{k}}<m \end{cases}. \tag{17}$$

Let $d(a,b,\mathbf{l})$ and $d^{(2)}(a,b,\mathbf{l})$ be the first two moments of $D(a,b,\mathbf{l})$.

We can now enumerate the events possible at the end of every slot, which will yield the desired sets of equations. Describing all the different possible events for the general case is quite tedious, though possible. Even for the two servers case, $m=2$, this description is lengthy, and we thus only provide the first few possible events.

In the case $m=2$ the state definition includes a, b, and l - the number of customers in the other queue. The possible events are:

1. A customer leaves the system from the local queue. The probability of this event is $\tilde{\mu}(a,b,l)$. If $a>0$ C's state will change to $(a-1,b,l)$. Otherwise C leaves the system

2. For $l=0$: A customer arrives at the system. The probability of this event is $\tilde{\lambda}(a,b,l)$. As $l=0$, the customer will always choose the other queue, and thus C's state will change to $(a,b,1)$.

And so on.

4) Numerical Results for the M/M/2 Model: We consider a system consisting of two servers. Fig. 6 depicts the results from evaluating the sets of equations provided in Section IV-A. The results are also supported by simulation. For all plotted points $\mu=1$ while λ changes according to ρ. The figure demonstrates the following properties:

1. For all systems, the unfairness is monotone increasing with the system utilization factor ρ.

2. Similarly to the situation in the M/D/2 model, the single queue system is more fair than all the multiple queue systems, for every system load. The difference in unfairness between this system and the best multiple server system, in our analysis the head-jockeying-on-idle system, is around 10%. Recall that as long as jockeying is allowed only in a manner that is nondiscriminatory with respect to the required service time (as is the case in the systems analyzed) the single queue system is at least as efficient as the multiple queue systems.

3. Similarly to the situation in the M/D/2 model, among the multiple queue systems, the jockeying-on-idle systems are more fair than the idling system, for every system load. The difference between these two types of systems is quite substantial. At lower loads, the ratio between them can be more than 1:10.

4. Similarly to the situation in the M/D/2 model, among the jockeying-on-idle systems, head-jockeying is more fair than tail-jockeying for every system load. The difference between them is around 10%. Recall that these two systems are equally efficient.

To summarize, all three properties discussed in Section III-B for the G/D/m model (and demonstrated to be true for the M/D/2 model in Section III-D) are demonstrated to be true for the M/M/2 model.

B. The Effect of Queue Joining Policy on Fairness

In this section we briefly analyze the queue joining policies discussed in Section III-C, for M/M/m type systems. The RAND policy was analyzed in Section IV-A.3, and thus we only analyze the SQ and RR policies (with head-jockeying-on-idle). Again, we only show some interesting details of the analysis for the case of $m = 2$ (two servers and two queues), where the rest of the analysis can be found in [21]

1) Fairness in the Join The Shortest Queue Policy: The analysis of the SQ policy is almost identical to that of the RAND policy given in Section IV-A.3. We point out the major differences.

One difference is that C always joins the shortest queue, and therefore (17) should be corrected. A second difference is that since customers always join the shortest queue, the events are slightly different. For example, if $l < a + b + 1$ C's state changes to $(a, b, l + 1)$. This leads to a slightly different set of equations.

2) Fairness in the Round-Robin Queue Joining Policy: Let r denote the index of the queue that the next arriving customer will join, $r = 1, \ldots, m$. The state (a, b, l, r) now captures all that is needed to predict the future of C.

The analysis is again nearly identical to that of the RAND policy given in Section IV-A.3, except that r needs to be added to the state definition in every place it is used. As in the SQ policy, (17) should be corrected, and the set of events/equations is slightly different.

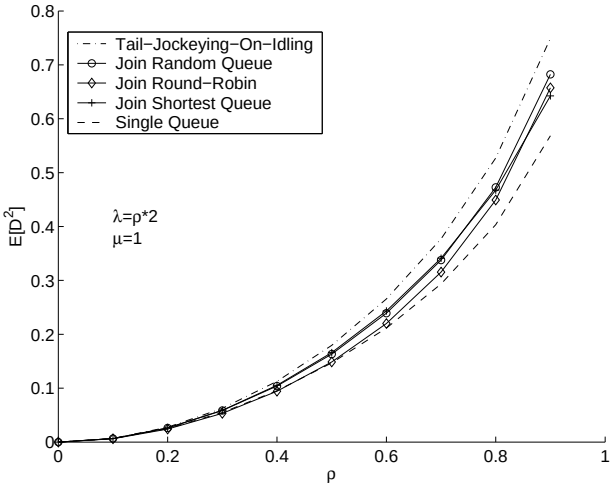

Fig. 7. The Effect of the Queue Joining Policy on the System Fairness

3) Numerical Results: Fig. 7 depicts the unfairness of the three joining policies analyzed - random, round-robin, and shortest queue, for the M/M/2 model, under head-jockeying-on-idle, as a function of the system utilization factor ρ. They were computed using the equations derived above. For comparison we also plot the results for the single queue system and the results for the tail-jockeying-on-idle system (with random queue joining). The results are also supported by simulation. The figure demonstrates the following properties:

1. All three queue joining policies are less fair than the single queue system, and more fair than the tail-jockeying system.

2. The unfairness of the system in the three queue joining policies is very similar. In fact, the unfairness is within a 1% interval, which might be caused by inaccuracies in our computation. This means that according to our analysis, in the case of head-jockeying-on-idle M/M/2 systems, the queue joining policy (among the three policies studied) has little effect on the fairness of the system.

Remark 4.1: The latter insensitivity to the queue joining policy can be explained by the fact that the jockeying alleviates potential discrimination caused by unfair queue joining policies.

V. Fairness of G/G/m Type Systems

In this section we consider $G/G/m$ type systems, i.e. systems where the customers arrive according to a general arrival process, the service requirements are generally distributed and the system has multiple servers. We show, via a simple example, that results similar to those derived in the analysis of the G/D/m system (Section III-B) and demonstrated to be true for M/M/m type systems (Section IV-A) do not necessarily hold for any arbitrary $G/G/m$ system. Nonetheless the question whether these results hold for $G/GI/m$ type systems (namely where the service times are generally distributed and *independent* of each other and of the arrivals), remains open and is left for future research.

In the following example we show that the single queue system is not necessarily more fair than a multiple queue system with no jockeying. Consider the following two server scenario, involving 4 customers, C_1, C_2, C_3 and C_4: $\{(a_i, s_i)\}_{i=1,2,3,4} = \{(0, s+1), (0, 2\epsilon), (\epsilon, s+1), (1, \epsilon)\}$, where $\epsilon \to 0$ and $s \gg \epsilon$ is some finite service requirement. If customers are served in a single queue, then C_4, which has a very short service requirement, has to wait s units of time to be served. As C_2 is served immediately upon arrival and its service time is very small, we can disregard this interval of service in our numerical calculation. This leads to a relatively high unfairness, namely $2 \times (s \times 1/3)^2 + (s \times (-2/3))^2 = 2s^2/3$. Consider now the two queues (q_a and q_b) case where jockeying is not allowed. Assume C_1 joins q_a and C_2 can join q_b. C_3 joins the system at epoch ϵ, finding both serves busy, and joins q_a (using round-robin joining policy, or randomly). C_4 can then be served upon arrival. As C_2 and C_4 are served immediately upon arrival and their service times are very small, we can disregard these intervals of service in our numerical calculation. As C_3 only needs to wait for one customer, namely C_1, compared to two customers for which C_4 needs to wait for in the previous case, the unfairness is only $(s \times 1/2)^2 + (s \times (-1/2))^2 = s^2/2 < 2s^2/3$.

This same example also shows that the standard (idling) system is not always less fair than the jockeying-on-idle systems, and similar scenarios can be constructed to show that other properties do not hold as well

Next we consider the properties proven in Section III-B for the G/D/m model and demonstrated in Section III-D for the $M/M/m$ model. While they do not hold for an arbitrary $G/G/m$ model we provide sufficient conditions under which these properties hold for the $G/GI/m$ model, i.e. where customer service requirements are i.i.d. random variables. We start with some definitions.

Define the *residual warranted service* of customer l at epoch τ, $R_l(\tau)$, as the integral of warranted service rate from epoch τ onwards, i.e. $R_l(\tau) = \int_\tau^{d_l} \omega(t)/N(t)dt$, $a_l < \tau < d_l$.

Define a (C_j, C_k)-*sample path* to be a set of arrival epochs for all customers, and service requirements for all customers, except a pair C_j, C_k. In the following we assume that the service time of C_i is a random variable denoted S_i (with realization s_i).

A system is said to be *Forward Dominant* if for every service policy $\phi \in \Phi$, for every pair C_j, C_k, then for every (C_j, C_k)-sample path for which $g_j < g_k$, the expected residual warranted service obeys $E[R_j(g_j)] < E[R_k(g_j)]$, where expectations are taken on the joint distribution of S_j, S_k.[3] For example, it is easy to see that if all customers have identical service times, the system is Forward Dominant, as is the case for single server systems. In fact in these cases it is true for every pair of service requirements s_j, s_k and not only in expectation.

[3]note that since we consider $\phi \in \Phi$, then the condition $g_j < g_k$ does not depend on the actual values of S_j and S_k, and thus taking the expectation over the full joint distribution of S_j and S_k is well defined.

Theorem 5.1 (Preference of Seniority for G/GI/m): Consider a Forward Dominant system. Let C_j and C_k be two customers, such that $a_j < a_k$ and their service requirements are i.i.d. random variables S_j, S_k, with realizations s_j, s_k. Let $\phi \in \Phi$ be a service policy. If there exists a (C_j, C_k)-sample path such that $g_k < g_j$, then there exists a service policy $\phi' \in \Phi$ such that $g_j < g_k$ whose expected unfairness (where expectations are taken over the joint distribution of S_j and S_k) is lower than that of ϕ.

Proof: First we construct ϕ' from ϕ: Until g_k ϕ' is identical to ϕ. At g_k it chooses to serve C_j (whose service time is s_j). From $g_k + s_j$ and on ϕ' imitates what ϕ would do in the case that $C_k = s_j$, with the exception that when ϕ serves C_j ϕ' serves C_k. Since $\phi \in \Phi$ and ϕ' imitates it, $\phi' \in \Phi$.

Next we compare the unfairness of ϕ and ϕ'. We do this for any sample path, and then take expectation over the service requirements of all customers. For $s_j = s_k$, ϕ' serves all customers in the same order as ϕ, except that it interchanges C_j and C_k. From Lemma 3.1 it follows that the unfairness of ϕ' is lower in this sample path.

For $s_j \neq s_k$ we need to track two sample paths for each policy: For every pair of service times τ_1, τ_2 (wlg assume $\tau_1 < \tau_2$) we consider one path where $S_j = \tau_1$ and $S_k = \tau_2$ (sample path (a)) and one path where $S_j = \tau_2$ and $S_k = \tau_1$ (sample path (b)) and apply both ϕ and ϕ' on both.

Let $F^{(a)}, F^{(b)}$ denote the unfairness for ϕ in the respective sample path, and similarly let $F^{(c)}, F^{(d)}$ denote the unfairness measured for ϕ' in the same paths. Note that the probability of $S_j = \tau_1, S_k = \tau_2$ is equal to that of $S_j = \tau_2, S_k = \tau_1$, since S_j and S_k are i.i.d. We can therefore show that the expected overall unfairness under ϕ is larger than that under ϕ', by showing that $F^{(a)} + F^{(b)} - F^{(c)} - F^{(d)} > 0$.

Let $\hat{F^{(a)}}$ denote the total unfairness under (a) due to C_j and C_k, and $\tilde{F^{(a)}}$ denote the total unfairness under (a) due to all customers other than C_j and C_k. Then $F^{(a)} = \tilde{F^{(a)}} + \hat{F^{(a)}}$. We define similar notations for (b), (c) and (d). Note that for any customer other than C_j and C_k, (a) and (c) are identical, and so are (b) and (d). Therefore $\tilde{F^{(a)}} = \tilde{F^{(c)}}$ and $\tilde{F^{(b)}} = \tilde{F^{(d)}}$. Thus $F^{(a)} + F^{(b)} - F^{(c)} - F^{(d)} = \hat{F^{(a)}} + \hat{F^{(b)}} - \hat{F^{(c)}} - \hat{F^{(d)}}$.

Let $R_{(1)}, R_{(2)}$ denote the warranted service in the intervals (a_j, a_k) and (a_k, g_k) respectively. These are for both paths for both policies. Let $R_j^a(g_k), R_k^a(g_k)$ denote the residual warranted service from epoch g_k, for customer C_j, C_k respectively, under ϕ for scenario (a). Denote similar variables for for scenario (b).

$$
\begin{aligned}
&\hat{F^{(a)}} + \hat{F^{(b)}} - \hat{F^{(c)}} - \hat{F^{(d)}} = \\
&\frac{1}{L}\big[\big(\tau_1 - (R_{(1)} + R_{(2)} + R_j^a(g_k))\big)^2 + \big(\tau_2 - (R_{(2)} + R_k^a(g_k))\big)^2 \\
&+ \big(\tau_2 - (R_{(1)} + R_{(2)} + R_j^b(g_k))\big)^2 + \big(\tau_1 - (R_{(2)} + R_k^b(g_k))\big)^2 \\
&- \big(\tau_2 - (R_{(1)} + R_{(2)} + R_k^a(g_k))\big)^2 - \big(\tau_1 - (R_{(2)} + R_j^a(g_k))\big)^2 \\
&- \big(\tau_1 - (R_{(1)} + R_{(2)} + R_k^b(g_k))\big)^2 - \big(\tau_2 - (R_{(2)} + R_j^b(g_k))^2\big]
\end{aligned}
$$

$$= \frac{2}{L} R_{(1)} (R_j^a(g_k) + R_j^b(g_k) - R_k^a(g_k) - R_k^b(g_k)). \quad (18)$$

We now take expectation of this expression over the joint distribution of S_j and S_k. Note that $R_{(1)}$ is constant in this context, thus the sign of the expected value of this expression is determined by the sign of the expected value of $R_j^a(g_k) + R_j^b(g_k) - R_k^a(g_k) - R_k^b(g_k)$, which is positive due to the system being forward dominant. ■

From this theorem it follows that if a G/GI/m system is forward dominant then the properties shown in Lemma 3.1 and Theorem 3.3 hold (with the modification that there the result holds deterministically, and here in expectation over the service time distribution).

VI. Practical Implications

The configurations studied in this work are used in many applications; the reader can easily identify them with his/her daily life encounters in airports, banks, Call Centers and Web servers. The designers of these systems, who care about the fairness of their system, can use our results in several ways. First, for the systems analyzed here the models and results can be used directly. For systems with general service times our models can be combined with those reported in [22] to carry out the analysis (as long as the dimensionality does not increase too much). For systems with larger number of queues our results can potentially be generalized, provided care is taken with regard to the large state space. Lastly, when the computational complexity becomes to large (due to large state space) one can easily use a simulation program to qualitatively evaluate the fairness level (via RAQFM). The importance of our results, with this respect is that they provide an analytic base to compare to, as well as an intuitive foundation that is required in order to build confidence in this newly proposed abstract measure.

VII. Concluding Remarks

Our work aimed at studying the fairness aspects of multi-queue and multi-server operational strategies and mechanisms. We used the RAQFM measure to quantitatively evaluate the system fairness under various common strategies. We applied our analysis to the G/D/m and M/M/m models. For the former model we showed that: 1) Global FCFS is the most fair scheduling, 2) The single-queue system is more fair than the multi-queue system, 3) Jockeying-on-idle from the head of the queue is more fair than from the tail of the queue, and 4) The Shortest-queue and the Round-Robin queue joining policies are more fair than the Random queue joining policy. For the latter model we provided an exact analysis of the system under the various operational strategies. We evaluated it numerically over a wide range of parameters for the M/M/2 case. The results demonstrated full support of the first three properties, and showed relative insensitivity to the queue joining policy. For the M/D/2 model and M/M/2 models we also demonstrated that jockeying-on-idle decreases the system unfairness. Simulation results, conducted on an M/GI/2 model, where the service time is of high variability, supported some of these properties as well.

The exact analysis of the M/GI/m model as well as several operational issues (e.g., arbitrary jockeying and queue dedication to short jobs) remain open for further research.

Acknowledgment

This work was supported in part by the Israeli Ministry of Science and Technology, grant number 380-801, and by EURO-NGI network of excellence.

References

[1] C. Palm, "Methods of judging the annoyance caused by congestion," *Tele. (English Ed.)*, vol. 2, pp. 1–20, 1953.

[2] I. Mann, "Queue culture: The waiting line as a social system," *Am. J. Sociol.*, vol. 75, pp. 340–354, 1969.

[3] W. Whitt, "The amount of overtaking in a network of queues," *Networks*, vol. 14, no. 3, pp. 411–426, 1984.

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

[5] R. C. Larson, "Perspective on queues: Social justice and the psychology of queueing," *Operations Research*, vol. 35, pp. 895–905, Nov-Dec 1987.

[6] A. Rafaeli, G. Barron, and K. Haber, "The effects of queue structure on attitudes," *Journal of Service Research*, vol. 5, no. 2, pp. 125–139, 2002.

[7] A. Rafaeli, E. Kedmi, D. Vashdi, and G. Barron, "Queues and fairness: A multiple study investigation," Faculty of Industrial Engineering and Management, Technion. Haifa, Israel. Under review, Tech. Rep., 2003. [Online]. Available: http://iew3.technion.ac.il/Home/Users/anatr/JAP-Fairness-Submission.pdf

[8] B. Avi-Itzhak, H. Levy, and D. Raz, "Quantifying fairness in queueing systems: Principles, approaches and applicability," RUTCOR, Rutgers University, Tech. Rep. RRR-25-2005, August 2005, submitted. [Online]. Available: http://rutcor.rutgers.edu/pub/rrr/reports2005/25_2005.pdf

[9] E. Sherry-Gordon, "New problems in queues: Social injustice and server production management," Ph.D. dissertation, MIT, May 1987.

[10] D. Raz, H. Levy, and B. Avi-Itzhak, "A resource-allocation queueing fairness measure," in *Proceedings of Sigmetrics 2004/Performance 2004 Joint Conference on Measurement and Modeling of Computer Systems*, New York, NY, June 2004, pp. 130–141, (*Performance Evaluation Review*, 32(1):130-141).

[11] ——, "RAQFM: A resource allocation queueing fairness measure," RUTCOR, Rutgers University, Tech. Rep. RRR-32-2004, September 2004. [Online]. Available: http://rutcor.rutgers.edu/pub/rrr/reports2004/32_2004.ps

[12] B. Avi-Itzhak and H. Levy, "On measuring fairness in queues," *Advances in Applied Probability*, vol. 36, no. 3, pp. 919–936, September 2004.

[13] D. R. Smith and W. Whitt, "Resource sharing efficiency in traffic systems," *Bell System Technical Journal*, vol. 60, pp. 39–55, 1981.

[14] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14, pp. 181–189, 1977.

[15] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," *IEEE transactions on Automatic Control*, vol. 25, pp. 690–693, 1980.

[16] A. Hordijk and G. Koole, "On the optimality of the generalized shortest queue policy," *Probability in the Engineering and Informational Sciences*, vol. 4, pp. 477–487, 1990.

[17] W. Whitt, "Deciding which queue to join: Some counterexamples," *Operations Research*, vol. 34, no. 1, pp. 55–62, Jan./Feb. 1986.

[18] D. Raz, B. Avi-Itzhak, and H. Levy, "Classes, priorities and fairness in queueing systems," RUTCOR, Rutgers University, Tech. Rep. RRR-21-2004, June 2004. [Online]. Available: http://rutcor.rutgers.edu/pub/rrr/reports2004/21_2004.pdf

[19] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. Wiley, 1975.

[20] R. Wolff, "Poisson arrivals see time averages," *Oper. Res.*, vol. 30, no. 2, pp. 223–231, 1982.

[21] D. Raz, B. Avi-Itzhak, and H. Levy, “Fairness considerations in multi-server and multi-queue systems,” RUTCOR, Rutgers University, Tech. Rep. RRR-11-2005, February 2005. [Online]. Available: http://rutcor.rutgers.edu/pub/rrr/reports2005/11_2005.pdf

[22] E. Brosh, H. Levy, and B. Avi-Itzhak, “The effect of service time variability on job scheduling fairness,” RUTCOR, Rutgers University, Tech. Rep. RRR-24-2005, July 2005. [Online]. Available: http://rutcor.rutgers.edu/pub/rrr/reports2005/24.pdf