Operations Research Perspectives 5 (2018) 145–149

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

The expected discrimination frequency for two-server queues

Berenice Anne Neumann[a], Hendrik Baumann*[,b]

[a] Research Group Statistics and Stochastic Processes, University of Hamburg, 20146 Hamburg, Germany
[b] Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, 38678 Clausthal-Zellerfeld, Germany

ARTICLE INFO

Keywords:
Queueing
Fairness measures
Multi-server queue
Combined queue vs. separate queues

2010 MSC:
60K25
68M20
60J28

ABSTRACT

Fairness measures for queues were introduced for measuring the individual satisfaction of human customers with respect to the waiting experience. The measure which performs best in some sense is the expected discrimination frequency (DF). In contrast to competing fairness measures, up to now, the DF has not been thoroughly analysed for multi-server systems. In particular, there are no results concerning the question whether or not in terms of the DF, combined queues are fairer than separate queues. In this note, we prove that under Markovian assumptions, combined queues are fairer and, furthermore, that this statement does not remain true for general queueing systems.

1. Introduction

Traditionally, the system performance of queueing systems is measured by characteristics such as waiting times, throughput, ... In recent years, fairness measures have been paid attention to. Considering fairness in queues has various reasons, and therefore, various kinds of fairness measures have been introduced.

In computer applications, it is a quite natural approach to consider the proportion of the response time of a job of size x to its size x. This quotient is referred to as the *slowdown*. For a queue with stochastic arrival process and stochastic service times, by considering stationary behaviour and taking the expectation, the (un)fairness of scheduling disciplines can be classified [1,2]. It turns out that the disciplines PS (processor sharing) and preemptive LCFS (last come, first served) can be regarded as some kind of fair with respect to the expected slowdown.

In many applications of queueing theory, human customers are involved (for example, supermarkets, waiting rooms at doctor's offices, check-in areas at airports, ...). Whereas slowdown-based considerations intend to find an abstract classification of fairness, for systems with human customers, psychological aspects become important: Human customers will judge the system by means of the 'perceived fairness'. Based on their satisfaction with their waiting experience, they will decide whether or not to revisit the facility providing the waiting system in the future (if they have a choice). Usually, human customers will not judge preemptive LCFS as a fair scheduling discipline, and hence, the slowdown-based classification of (un)fairness cannot be applied in this context.

Pychological studies [3] revealed that human customers perceive 'unfairness' if they are overtaken by other customers or if customers with a larger job size are allowed to leave the system earlier. Based on these findings, principles for measuring perceived fairness have been established [4]: For single-server queues, fairness measures should fulfill a *seniority preference principle* and a *service-requirement preference principle*. In their strong version, tests for these principles require that

- if two jobs have the same service requirement, the job which arrived earlier should be completed first,
- if two jobs arrive at the same time, the job with smaller service requirement should be completed first.

In both cases, interchanging the order of service of the two jobs under consideration should lead to a lower fairness/ higher unfairness. In order to analyse perceived fairness, order fairness [5], a slowdown-based measure [6], the measure RAQFM (resource allocation queueing fairness measure) [7] and the discrimination frequency (DF) [8] have been introduced, further analysis can be found in [9–13]. In some way, the DF performs best with respect to the principles established in [4], since it is the only measure introduced so far which satisfies the strong tests both for the seniority principle and the service requirement principle.

For multi-server systems, there is psychological evidence that human customers generally judge single-queue systems fairer than multi-queue systems, see [3]. A measure being appropriate for evaluating the fairness of multi-server and multi-queue systems should

* Corresponding author.
E-mail addresses: berenice.neumann@uni-hamburg.de (B.A. Neumann), hendrik.baumann@tu-clausthal.de (H. Baumann).

https://doi.org/10.1016/j.orp.2018.06.001
Received 8 January 2018; Received in revised form 27 April 2018; Accepted 12 June 2018
Available online 19 June 2018

reflect this judgement. For the RAQFM, an analysis has been performed in [10], yielding that for $G/D/k$ and $M/M/2$ models, the single-queue is fairer than the multi-queue. However, it is shown that this result does not hold for general $G/G/k$ queues. The main goal of this paper is to provide a similar analysis for the DF in the case of simple Markovian systems. We will focus on the FCFS discipline, but nevertheless, our results can be interpreted as a starting point for a future investigation of the impact of the scheduling discipline on the discrimination frequency in multi-server systems.

The structure will be as follows: In Section 2, we will describe the considered single-queue and multi-queue system, and restate the precise definition of the discrimination frequency. In Section 3, we will derive the expected DF for the single-queue system, and in Section 4, we will determine a lower bound for the expected discrimination frequency in the multi-queue system and prove that indeed, in terms of the DF, the single-queue system is fairer than the multi-queue system. In Section 5, we will present an example that for general (non-Markovian) systems this statement does not remain true. In Section 6, we will summarize our results, and we will outline possible directions of further research.

2. Basic terms and models under consideration

In this paper, we aim for comparing the expected discrimination frequency for an $M/M/2$-model and two $M/M/1$ models with separate queues. We briefly present both models and the precise definition of the discrimination frequency.

2.1. The M/M/2 model

Customers arrive according to a Poisson process with intensity λ. There are two identical servers, and the service times are independently and identically Exp(μ)-distributed. Furthermore, there is no restriction of the number of waiting customers, and the scheduling discipline is FCFS (first come, first served). Due to these modelling assumptions, the process $(N_t)_{t\geq 0}$ of the number N_t of customers in the system (waiting in the queue or being served) is a continuous-time Markov chain (CTMC). In case $\rho = \frac{\lambda}{2\mu} < 1$, the system is stable, and in the long-run, it will behave stationarily, that is, for any $k = 0, 1, 2, \ldots$, we have $\lim_{t\to\infty} \mathbb{P}(N_t = k) = \pi_k$, where $\pi = (\pi_k)_{k=0}^{\infty}$ is the stationary distribution. It is well-known [14, Section 3.5] that

$$\pi_0 = \frac{1-\rho}{1+\rho} \qquad \text{and} \qquad \pi_n = 2\rho^n\pi_0, \ \ n \geq 1.$$

Due to the PASTA property of the arrival process [15, Theorem VII.6.7], in the long-run, arriving customers will 'see' the stationary distribution, that is, with probability π_k, an arriving customer will find k other customers in the system. Note that for the stationary number N of customers in the system, we have $\mathbb{E}[N] = \sum_{n=0}^{\infty} n\pi_n = \frac{2\rho}{(1-\rho)(1+\rho)}$.

2.2. The multi-queue model

In order to model two separate queues, we consider two parallel $M/M/1$ models. Customers still arrive according to a Poisson process with parameter λ. Each arriving customer will join the first system with probability $\frac{1}{2}$, and the second one with probability $\frac{1}{2}$. Hence, the arrival process for each of both systems is a Poisson process with intensity $\frac{\lambda}{2}$. Both systems have one server, and the service times are independently and identically Exp(μ) distributed. Still, we assume infinite waiting capacity and FCFS as scheduling discipline. Let $(N_t^{(1)}, N_t^{(2)})_{t\geq 0}$ be the process of the number of customers in the first and in the second system, respectively. Due to the modelling assumptions, this process is again a CTMC, and furthermore, $(N_t^{(1)})_{t\geq 0}$ and $(N_t^{(2)})_{t\geq 0}$ are independent, and both are CTMCs. For $\rho = \frac{\lambda}{2\mu}$, we have stability, and

$$\lim_{t\to\infty} \mathbb{P}(N_t^{(1)} = k) = \lim_{t\to\infty} \mathbb{P}(N_t^{(2)} = k) = \pi_k, \qquad k = 0, 1, 2, \ldots,$$

where $\pi = (\pi_k)_{k=0}^{\infty}$ is the stationary distribution. Again, the exact shape of π is well-known [14, Section 3.2], we have $\pi_k = (1-\rho)\rho^k$ for all $k = 0, 1, 2, \ldots$. As for the $M/M/2$ model, we have the PASTA property, that is, in the long-run, with probability $\pi_k \cdot \pi_\ell$ an arriving customer sees k other customer in the first system, and ℓ other customers in the second system.

Although we will compare the fairness (measured by the discrimination frequency), we briefly recapitulate that traditional performance measure favor the combined queue over the separate queue: Let N be the total stationary number of customers in the system. Then $N = N^{(1)} + N^{(2)}$ and $\mathbb{E}[N] = \frac{2\rho}{1-\rho}$, and this number is larger (by factor $1+\rho$) than the corresponding expected number of customers in the $M/M/2$ queue. Since Little's formula guarantees that the expected response (or sojourn) time of any 'black box' can be determined by $\frac{\mathbb{E}[N]}{\lambda}$, this result carries over to response times.

Note that there are different ways to 'choose' the queue an arriving customer joins. Here, we consider the 'coin toss'. A natural alternative is joining the shorter queue (if there is one). In this case, the stationary numbers of customers in the systems depend on each other. We leave this topic open for future research.

2.3. The discrimination frequency

The discrimination frequency was introduced in [8]. The intuitive concept behind it is to count the discriminating events a customer suffers from. These are *large jobs*, that are jobs which have a larger remaining service requirement at our job's time of arrival, but leave the system earlier, and *overtakes*, that are jobs which arrive after and leave before our marked job. Formally, in [8], the DF was defined as follows:

Definition 2.1. Let a_i, d_i, s_i be the arrival time, the departure time, and the service time of job J_i. Furthermore, let $s_j'(t)$ be the residual service time of J_j at time t (if J_j did not enter the system at time t, we have $s_j'(t) = s_j$). Then the amount $OV(i)$ of overtakes job J_i suffers from is

$$OV(i) := |\{j: (a_j \geq a_i \wedge d_j \leq d_i)\}|.$$

The amount $LJ(i)$ of large jobs that a job J_i suffers from is

$$LJ(i) := |\{j: (d_i \geq d_j > a_i \wedge s_j'(a_i) \geq s_i\}|.$$

The discrimination frequency of job J_i is

$$DF(i) = OV(i) + LJ(i).$$

The discrimination frequency of a system in steady state is the discrimination frequency of a stationary customer.

For stationary systems, the distribution of $OV(i)$, $LJ(i)$, and $DF(i)$ is identical for all customers i. We will refer to the number of overtakes, the number of large jobs, and the discrimination frequency of a randomly chosen customer as OV, LJ, and DF respectively. Hence, we will consider a 'tagged customer' who sees the stationary distribution of number of customers in the instant of his arrival, and we will pursue his way through the system, and count the number of overtakes and large jobs he suffers from.

3. The expected discrimination frequency for the combined queue

In order to compute $\mathbb{E}[DF]$, we determine $\mathbb{E}[LJ]$ and $\mathbb{E}[OV]$. Note that under FCFS, large jobs are only caused by customers which have entered the system before our tagged customer, and overtakes are only caused by customers which will enter the system after our tagged customer. Precisely, we will prove the following result in the next subsections.

Theorem 3.1. *For an $M/M/2$ model with a combined FCFS, the expected number of large jobs is $\mathbb{E}[LJ] = \frac{\rho^2}{(1-\rho)(1+\rho)}$, the expected number of overtakes in the $M/M/2$-system with a combined FCFS queue is $\mathbb{E}[OV] = \frac{\rho}{1+\rho}$, and the expected discrimination frequency is*