Operations Research Perspectives 5 (2018) 133–144

Contents lists available at ScienceDirect

Operations Research Perspectives

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

# On slowdown variance as a measure of fairness

Maccio Vincent J., Hogg Jenell, Down Douglas G.*

*Department of Computing and Software, McMaster University, Hamilton, Ontario L8S 4L7, Canada*

A B S T R A C T

When considering fairness one must ask two fundamental questions. Firstly, what does it mean to be fair? And secondly, how does one measure that fairness? Different authors have offered different notions and metrics to address these questions. We provide arguments identifying where past metrics fall short, discuss how the underlying motivations differ, and offer our own metric to address these issues. That is, we propose using a system's *slowdown variance* (SDV) as a measure for its fairness. Advantages of SDV are demonstrated via a suite of simulation experiments which compare a range of established policies under a range of service time distributions. These advantages include a decoupling of fairness from performance, an intuitive distinction between *last come first serve* and *processor sharing*, as well as recognition of starvation within *shortest remaining processing time*.

## 1. Introduction

Fairness in queueing systems has been an active topic in the stochastic modelling research community [19]. Perhaps the most fundamental questions in this domain are "What does it mean to be fair?" and "How can fairness be measured?". In the literature this is still a point of debate. That is, there does not exist a universally agreed upon definition of fairness. Moreover, different definitions of fairness may be in direct conflict or be of questionable applicability in certain contexts. We do not aim to determine which previously defined metrics are better than others. Rather, we offer a notion of fairness which is motivated by past definitions while looking to address some of the concerns these past metrics raise. Specifically, we propose using the slowdown variance (SDV) as a metric for fairness.

In order to understand the motivations behind our proposal, one must first understand how fairness has been presented in the research community. There are two fundamental views of fairness to consider: *temporal fairness* (also referred to as seniority-based fairness) and *proportional fairness*. Temporal fairness suggests that jobs should be served in order of their seniority. That is, if job $J_1$ arrives before job $J_2$, then job $J_1$ should be completed before $J_2$. A variety of fairness metrics such as politeness and order fairness provide means to measure a policy's temporal fairness. Our work does not discuss temporal fairness or its measures in any level of detail. For further reading, we direct the interested reader to Wierman [18] and Avi-Itzhak and Levy [3]. Instead, our work focuses on, and is inspired by, the notion of proportional fairness. Proportional fairness posits that a job's response time should be proportional to its size. That is, informally, if a small job arrives shortly after a large job, it may be fair for the small job to complete before the large job does. In other words, the longer a job is going to take to serve the more fair it is that it waits in queue. For a comprehensive discussion and comparison of metrics pertaining to both temporal and proportional fairness, we direct the reader to Avi-itzhak et al. [5].

Popular metrics for evaluating fairness are often times measures of the system's slowdown [6,7,10–12,17,20,21]. Here, slowdown is defined as $R/S$, where $R$ is the response time of a job and $S$ is its size or service time (this work uses the two interchangeably for ease of exposition). The justification of this metric stems from the proportionality principle; it is fair that a job has a response time proportional to its service time (larger jobs should wait longer). It is worth noting that while researchers look to the slowdown as a base for fairness, slowdown is also considered a measure of performance; the response time is present in the numerator.

With regards to fairness, one of the first measures based off slowdown was simply examining the maximum slowdown [7]. This gives insight into the worst-case scenario and in turn bounds it in many cases, but does not give a picture of how fair a system is in the average case. To achieve a better overall understanding, authors began looking at forms of the slowdown's expectation, i.e. $\mathbb{E}[R/S]$. The *conditional expected slowdown* [6] was introduced in response to growing interest in size-based policies that were suspected of treating larger jobs unfairly. It is defined as $\mathbb{E}[R(x)/x]$, where $R(x)$ is the expected response time for a job of size $x$. This metric provides finer-grained information than the expected slowdown. For example, it allows one to determine if certain job sizes are treated poorly.

* Corresponding author.
*E-mail addresses:* macciov@mcmaster.ca (V.J. Maccio), hoggjm@mcmaster.ca (J. Hogg), downd@mcmaster.ca (D.G. Down).

https://doi.org/10.1016/j.orp.2018.05.001
Received 6 January 2018; Received in revised form 23 May 2018; Accepted 23 May 2018
Available online 31 May 2018

Wierman and Harchol-Balter [20] leveraged the conditional slowdown to sort policies into categories. Based on the intuition that processor sharing (PS) is an inherently fair policy, they judge whether a policy is "Always Fair", "Sometimes Fair", or "Always Unfair". If a policy achieves a conditional expected slowdown that is less than or equal to the conditional expected slowdown offered by PS for all job sizes under any distribution it is Always Fair. If the conditional expected slowdown is less than or equal to that of PS for some job sizes under some distributions (but not all) it is Sometimes Fair. However, we find that this metric and criteria for evaluating proportional fairness have drawbacks. For one, the analysis in [20] is confined to an $M/G/1$ setting and the authors comment that it is unknown whether similar criteria can be derived for systems with general arrival processes and/or with multiple servers. Also, the categorization of policies into Always Fair, Sometimes Fair, and Always Unfair does not allow one to compare policies that fall into the same category, resulting in a partial ordering among policies.

In [13] Raz et al. proposed the Resource Allocation Queueing Fairness Measure (RAQFM), which was studied further by Avi-Izthak et al. in [4]. RAQFM is based off a *discrimination* measure, which in essence determines how far a policy is from giving each job an equal portion of the resource at every point in time. That is, it measures how much a policy differs from PS. This gives means to determine a total ordering of policies. However, this measure of fairness is still tightly coupled to the assumption that PS is perfectly fair. This assumption seems intuitive on the surface, but as will be seen later on in this work, can at times be problematic. Around the same time, Avi-Izthak et al. [2] also introduced the Slowdown Queueing Fairness measure (SQF). SQF, alongside its distinction from the SDV, is discussed in more detail in Section 2. In [2], it was proposed that a fair policy should offer every job a constant slowdown. The authors argued that the expected slowdown may not be sufficiently sensitive for evaluating fairness since policies can have similar expected slowdowns but have different behaviours about the expectation. The idea is that one can measure fairness with reference to an "ideally fair" slowdown.

Another means to measure fairness, the *discrimination frequency*, was introduced by Sandmann in [15] and discussed within the wider fairness framework in [16]. Here, a discrimination is counted every time one of the following events occurs: 1) a job arrives after another but completes before it, and 2) a job with greater remaining service time upon another job's arrival departs before the arriving job. Intuitively, the less frequently discriminations occur, the more fair the system is. Note, as is often the case with fairness, these two types of discrimination are in contention with each other. While the discrimination frequency is most certainly an attractive means to capture the tension between competing fairness properties, it belongs to a different school of thought than slowdown-based metrics (which is our focus).

As one can surmise there are several philosophies when measuring fairness. Moreover, fairness can be extremely sensitive to context. For example, what one deems to be fair for clients on a web-server, may be drastically different compared to what one deems to be fair for customers in a grocery store checkout. One notable difference between these two contexts is that in the grocery store checkout customers are free to see when other customers arrive, when they leave, how many items they are purchasing, etc. (white-box service), while on a web-server clients are often oblivious to how they are served relative to others (black-box service). This white-box/black-box differentiation can be seen in the fairness metrics as well. If for a given metric two policies with identical arrival and departure times for all jobs can be evaluated as having different levels of fairness then that metric is a white-box metric, otherwise, it is a black-box metric. Metrics like discrimination frequency and RAQFM are concerned with the details of how jobs are being served – how often jobs are being over taken, at a given point in time do all jobs have an equal share of the processor, etc; they are white-box metrics. On the other hand, metrics such as SQF, expected slowdown, and conditional slowdown are only concerned with the end performance, and have no regard as to how that performance is achieved; they are black-box metrics.

Our work's motivations are consistent with [2], but we offer a complementary viewpoint. We suggest that a truly fair policy would ensure that the ratio between a job's size and its response time remains constant (or as close to constant as possible). Therefore, the closer the SDV is to zero, the more fair a policy is. As observed in [2], we shall see that using this basis for fairness will provide insights that using expected or conditional slowdown would otherwise not capture. Most notably, using the expected slowdown or the conditional slowdown is known to equate two well-known policies, processor sharing and last come first serve, in terms of their fairness. However, we find evidence that SDV will determine one to be more fair than the other, as will be discussed in Section 3. Furthermore, since SDV is a black-box metric, for consistency of context and application this work focuses on comparing it against other black-box metrics.

Any mention of SDV in the literature has been brief – to our knowledge no such study or presentation of SDV exists. One possible explanation for this absence is that analytic expressions for SDV are difficult to determine (this appears to be a fundamental problem, as we have made attempts to generate analytic results). As such, we sacrifice analytic tractability in order to examine these ideals of fairness empirically. The contributions of this paper include, but are not limited to,

1. The introduction and justification of using SDV as a fairness metric, found in Section 2.
2. An extensive simulation study pertaining to SDV and expected slowdown under different scheduling policies and distributions, found in Section 3.
3. Several key observations and insights into the behaviour of SDV across these different configurations, also found in Section 3, which enriches the overall discussion of fairness.

## 2. Definitions and justification

As stated previously when discussing fairness, there are two important aspects which must be made clear. Firstly, what does it mean to be fair? And secondly, how does one quantify fairness? As seen in Section 1 these are subjective issues. Nevertheless, we believe metrics for fairness exist which are grounded in intuition, and moreover, such metrics are independent/decoupled from performance.

We proceed by addressing the broad but fundamental question of what it means to be fair. Consider a scenario where each customer $C$ has exact knowledge of its size, or service time, denoted by $S_C$, as well as its response time, denoted by $R_C$, but no knowledge of *how* it is served. Suppose that customer $C_1$ has a service time of one minute, $S_{C_1} = 1$, and a response time of three minutes, $R_{C_1} = 3$. With no information about other customers, they may simply perceive this as *typical* system performance. When a second customer $C_2$ is introduced, things become more complicated. Continuing with the example, let $S_{C_2} = 2$ and $R_{C_2} = 2$, so that $S_{C_1} < S_{C_2}$, but $R_{C_1} > R_{C_2}$. Therefore, we argue that even an impartial onlooker would view $C_1$ as being treated unfairly, as it requires less system capacity yet has a longer response time. As such, $C_1$ is likely to be dissatisfied. This dissatisfaction stems from $C_1$'s expectation of treatment relative to others. That is, it stems from $C_1$'s notion of fairness.

In this small example it is intuitive that all parties would agree the system is not fair. However, when *is* the system fair? A perfectly fair system can be demonstrated by considering the previous example where the response times vary according to the scheduling policy implemented. Assume the previous example was implementing the policy denoted by $\pi_1$. That is, under $\pi_1$, $R_{C_1} = 3$ and $R_{C_2} = 2$. Under another policy, denoted by $\pi_2$, suppose that $R_{C_1} = 1$ and $R_{C_2} = 3$. Lastly, under policy three, denoted by $\pi_3$, suppose that $R_{C_1} = 1.5$ and $R_{C_2} = 3$. Assuming both customers arrive at the same instant, a realization of these three policies is illustrated in Fig. 1, and the values are