Locality of Reference and the Use of Sojourn Time Variance for Measuring Queue Unfairness Extended Abstract

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

Benjamin Avi-Itzhak
RUTCOR, Rutgers University, New Brunswick, NJ, USA
aviitzha@rutcor.rutgers.edu

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

ABSTRACT

The variance of customer sojourn time (or waiting time) is used, either explicitly or implicitly, as an indication of fairness for as long as queueing theory exists. In this work we demonstrate that this quantity has a disadvantage as a fairness measure, since it is not local to the busy period in which it is measured. It therefore may account for customer discrepancies which are not relevant to fairness of scheduling. We show that RAQFM, a recently proposed job fairness measure, does possess such a locality property. We further show that within a large class of fairness measures RAQFM is unique in possessing this property.

1. INTRODUCTION

How should the fairness of a queueing system be quantified? Perhaps a very natural and appealing choice for a fairness measure is the variance of the sojourn time (or of the waiting time) as it measures the inequity in the delay suffered by the jobs in the system.

The waiting time variance was used as a measure of system unfairness as early as [2], where it is shown that among all non-idling policies where the customers are indistinguishable, First-Come-First-Served (FCFS) has the lowest variance and therefore is "in a sense the 'fairest' queue discipline". In [1] the waiting time variance was shown to be a surrogate measure for evaluating the order-displacement in the queue as well as as a metrics that measures the deviation of waiting times between customers. This also supports its use as a measure for queue unfairness. The question of why waiting time variance should not serve as a simple queue fairness measure was also posed to the authors in some personal communications and some referee reports. Lately, [4] use the scaled conditional variance of response time as a metric and criterion of "predictability",

We start this work by presenting a simple example exposing some weakness in using the sojourn time variance as a queue fairness metrics. To this end consider a single-server system where the service time is deterministic of 1 unit and arrivals occur in bulks consisting of either 2 in a bulk (type A) or 4 in a bulk (type B), with equal probability of either bulk. Assume that arrivals occur such that the inter-arrival times are always larger than 4 units, and thus each busy period consists of exactly one arrival. Now suppose that the server processes the jobs in a Processor Sharing (PS) mode. The sojourn time of all jobs are therefore 2 units and 4 units for type A and type B, respectively. The mean sojourn time is $3\frac{1}{3}$ and the sojourn time variance is $\frac{8}{9}$. This reflects significant variability of sojourn time which is an indication of unfairness. Nonetheless, an examination of the system reveals that all the jobs that are present concurrently in the system receive exactly the same treatment and thus there is no discrimination in the system. That is, the system is *fully fair*, which contradicts the measure of the sojourn time variance as an unfairness measure.

A careful examination of the system reveals the source for the difficulty in this example: The system variance accounts for inequity between the treatment of the type A and type B jobs (namely jobs in different busy periods). Nonetheless, from fairness point of view, each job cares only about other jobs that can affect its service. Thus, since a type A job cannot be affected by the service given to type B jobs (they are in different busy periods) it should not be compared to them. The use of the variance of sojourn time, therefore, yields a difficulty, since it involves a comparison of jobs that are not related to each other.

A second weakness in this approach is exposed by the following example. Consider the same system as above with the same arrival pattern. Now suppose that the server processes the jobs in a FCFS manner. The mean sojourn time is 2.167. Now focus on the job served second in a type A busy period. Its sojourn time is 2 compared to the mean sojourn time of 2.167, so this job receives better service than the mean, and can be considered "positively discriminated".

*This work was supported in part by grant 380-801 from the Israeli Ministry of Science and Technology

However, closer examination reveals that this is not the case - this job is negatively discriminated as it is served behind the only other job in the busy period. This reveals a second source of difficulty: A performance metric such as the sojourn time does not provide enough information to determine whether a specific job is positively or negatively discriminated.

2. LOCALITY OF REFERENCE AND COMPARISON SET

We next address the problems posed above. It does seem natural to use the variance of a performance metric (e.g. the sojourn time) as an indication of service inequity, that is of queue unfairness. Nonetheless, as demonstrate above, it is highly critical over which population such variance will be taken. The set of jobs over which the variance is taken is denoted *the comparison set*. The first question we address, therefore, is what comparison set should be used for evaluating fairness among jobs. Intuitively speaking, the example demonstrates that the performance of a job should be compared only to that of jobs that are time-wise "close to it", which we call "locality of reference".

More precisely, for a single-server work-conserving system (with no idling) we claim that the performance measures of two jobs should be compared to each other (that is, the jobs should be in the same comparison set) if and only if the two jobs are served in the same busy period. The reason is that the server can shift processing resources from one job to another if and only if the two jobs are processed in the same busy period.

The latter is established in the following theorem:

Theorem 2.1. *Consider two arbitrary jobs J_a and J_b. The server can transfer resources between them if and only if they reside in the same busy period.*

The notion of resource transferring is defined in detail in the full text of this paper. In broad terms, we define resource transferring as the existence of an alternative assignment of service processing times in which a period exists for which service once given to J_a is now given to J_b and vice versa, and the rest of the customers are not affected. We prove the theorem by showing how such an alternative service assignment can be built.

3. LOCALITY OF MEASUREMENT, LOCALITY OF VARIANCE AND THE RELATION BETWEEN THEM

Having proposed to use the variance of a performance metric as an indication of unfairness, and having realized that for the sake of fairness evaluation it is desired to have the comparison set consisting of the busy period, we now define two desire properties in such a measure designed to address the two difficulties raised in Section 1.

Assume that X is a random variable denoting a measure of an individual job (say waiting time, or sojourn time). Let Y be a random variable denoting the mean of the measure X in an arbitrary busy period under steady state.

Property 3.1 (Locality of Measurement). *A measure is said to be* Locally Measured *if for every service policy, Y is constant (not necessarily the same constant for all service policies).*

Intuitively speaking, if a metric is locally measured it provides a simple manner of determining whether a specific job is positively or negatively discriminated, by comparing its performance to the above mentioned constant (for the service policy used in the system).

For the second property we first define the concepts of *inter-variance* and *intra-variance*. The intra-variance of X is its second moment around the mean of X *at the same busy period*. This is in contrast to the regular variance, which we denote *global variance* in which the second moment is computed around the mean of X over the whole job population. In somewhat more formal terms, if J is a job and X is a random variable of a measure of the job (say, waiting time), and bp is a specific busy period (characterized by the number of jobs, their relative arrival times and service times) then the local variance is computed by first computing $E[X^2|J \in bp] - (E[X|J \in bp])^2$ and then unconditioning on bp. The inter-variance is defined as the second moment of Y around the mean of X.

Intuitively speaking, the intra-variance measures the diversification of specific measures from their mean in each busy period, while the inter-variance measures the diversification of busy period means.

The connection between the global variance, the inter-variance and the intra-variance is established in the following theorem:

Theorem 3.1. *The global variance equals the sum of inter-variance and intra-variance.*

proof sketch. If the number of jobs is given, this can be shown to be true for any grouping of jobs, and specifically to jobs in busy periods. From ergodicity arguments it follows that it is true in general. □

We can now define the second property:

Property 3.2 (Locality of Variance). *A metric is said to have* local variance *if its global variance equals its intra-variance for every service policy.*

The two properties are strongly related:

Theorem 3.2. *A metric is locally measured if and only if it has local variance.*

Proof. If a metric is locally measured its mean is constant for every busy period and therefore its inter-variance is zero. From Theorem 3.1 follows that it has local variance.

On the other hand, if a metric has local variance it follows that it has zero intra-variance, and therefore it is locally measured as well. □

4. LOCALLY MEASURED METRICS

We now investigate for which individual job measures the properties defined above hold.

We define a class of measures ξ which is relatively large. For a job J_i let a_i and d_i be its arrival and departure epochs respectively and let s_i be the amount of service the job was given. Let $N(t)$ be number of jobs in the system at time t. Each measure is identified by two functions, the warranted service function $f(N)$ and the utility function $g(S)$, the individual job measure being

$$X_i = g(s_i) - \int_{a_i}^{d_i} f(N(t))dt. \quad (1)$$

Specifically, we examine the *discrimination function* proposed in [3]. This individual job function aims at evaluating the deviation of the service a job receives from what it "deserves" to receive (and thus evaluates individual discrimination) and for job J_i is given by:

$$D_i = s_i - \int_{a_i}^{d_i} \frac{dt}{N(t)}. \quad (2)$$

In [3] it was proposed to use the variance of the random variable D (under steady state) as a measure of system unfairness. It is easy to see that $D \in \xi$, and so do the sojourn time and the waiting time.

Theorem 4.1. *Both Property 3.1 and Property 3.2 hold for the discrimination function D. This holds also for a class of functions which are very closely related to D, which we denote $\mathcal{D}$.*

Proof. It is easy to see that momentarily, the sum of the discriminations of all customers in the system at epoch t is zero. Thus, the average of discrimination over every period is zero, including the average over a busy period. Thus Property 3.1 holds and from Theorem 3.2 so does Property 3.2. □

We further show uniqueness of the above property, that is:

Theorem 4.2. *If either Property 3.1 or Property 3.2 holds for $X \in \xi$, then $X \in \mathcal{D}$.*

Proof Sketch. We focus on Property 3.1 since it was shown in Theorem 3.2 to be equivalent to Property 3.2. Note that Property 3.1 must hold for *every* service policy. We show that if it holds for X when the service policy is Processor Sharing, then $X \in \mathcal{D}$. Furthermore, as the definition of the properties can be weakened to hold just for non-preemptive policies, we show that if it holds for X for *any* non preemptive service policy, then $X \in \mathcal{D}$. Both proofs are done by building arrival patterns in which the mean of the measure in each busy period can be controlled. □

5. CONCLUSION

We show in this work that there are two main difficulties with using the variance of sojourn time as a measure for fairness, the first being that it involves comparisons between customers which are not related, and second being that the individual job measures lack a global reference.

We then show that RAQFM (the global variance of the discrimination D) does not have these difficulties, and is unique in this property within a large class of measures. Thus, one may conclude that using the (global) variance of D is appropriate for quantifying queue unfairness, while using the (global) variance of other queue measures (such as the variance of sojourn time) may sometimes lead to difficulties.

We should also note that the (analytic) derivation of the global variance of a job measure is typically much simpler than that of the local variance. For RAQFM they are the same, in which case the problem is avoided.

It is interesting to study whether the class of measures considered, ξ, can be made more general, making the statements in this work stronger. We suspect it can be done easily.

We would also like to point out that it might be possible to build specific measures that apply to smaller classes of service policies than the ones we considered, as long as they do not include processor sharing, or any non-preemptive policy. However, we believe that such limitations will make those measures impractical.

6. REFERENCES

[1] B. Avi-Itzhak and H. Levy. On measuring fairness in queues. *Advances in Applied Probability*, 36(3):919–936, September 2004.

[2] J. F. C. Kingman. The effect of queue discipline on waiting time variance. *Proceedings of the Cambridge Philosophical Society*, 58:163–164, 1962.

[3] 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*, pages 130–141, New York, NY, June 2004. (*Performance Evaluation Review*, 32(1):130-141).

[4] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to higher moments of conditional response time. In *Proceedings of ACM Sigmetrics 2005 Conference on Measurement and Modeling of Computer Systems*, pages 229–239, Banff, Alberta, Canada, June 2005.