RUTCOR
RESEARCH
REPORT

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

David Raz [a] Benjamin Avi-Itzhak [b]
Hanoch Levy [c]

RRR 35-2005, November, 2005

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, davidraz@post.tau.ac.il
[b]RUTCOR, Rutgers, the State University of New Jersey, 640 Bartholomew Road, Piscataway, NJ 08854-8003, USA, aviitzha@rutcor.rutgers.edu
[c]School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, hanoch@cs.tau.ac.il

Rutcor Research Report
RRR 35-2005, November, 2005

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

David Raz Benjamin Avi-Itzhak Hanoch Levy

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

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

1 Introduction

How should the unfairness of a queueing system be quantified? Perhaps a very natural and appealing choice for an unfairness metric 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 implicitly used as a metric of system unfairness as early as [3], where it is shown that among all non-idling policies where the jobs 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 metric for evaluating the order-displacement in the queue as well as a metric that measures the deviation of waiting times between jobs. This also supports its use as a metric for queue unfairness. The question of why waiting time variance should not serve as a simple queue unfairness metric was also posed to the authors in some personal communications and some referee reports. Lately, [5] uses the scaled conditional variance of response time as a metric and a criterion for evaluating queue "predictability".

To demonstrate the issue of this work let us examine the following simple example exposing some weakness in using the sojourn time variance as a queue unfairness metric. 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 for 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 average 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 supplied by the sojourn time variance as an unfairness metric.

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 unfairness 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 for the sake of evaluating unfairness. The use of the variance of sojourn time, therefore, yields some inaccuracy, 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 average 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 average sojourn time of 2.167, so this job receives better service than the average, and can be considered "positively discriminated". 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.

We start (Section 2) with defining the model and some notation. We then (Section 3) start addressing the difficulties presented above by introducing the concept of "Locality of Reference", and showing that there is a good reason to compare each job only to other jobs in the same busy period. We then (Section 4) define two desired properties, "Locality of Measurement" and "Locality of Variance", designed to dealing with the difficulties presented above. In Section 5 we investigate a rather large class of performance metrics and show that within this class of performance metrics only a small subclass of performance metrics, all related to the recently proposed fairness metric RAQFM, have these desired properties; neither the sojourn-time variance, nor the waiting-time variance belong to this subclass, and thus suffer from those difficulties. In Section 6 we consider a second approach one can use for devising a fairness metric, namely explicit evaluation of the intra-variance. In Section 7 we bring some numeric results, and concluding remarks are given in Section 8.

2 Model and Notation

We consider a single server system consisting of a server and a queue. Jobs, denoted $\{J_i\}_{i=1}^{\infty}$ are arriving at the system at arbitrary times according to the order $J_1, J_2, \ldots$. The arrival epoch of J_i is denote a_i. Each job has an arbitrary service requirement, denoted s_i. At each epoch some of the jobs in the system receive service by the server, according to the service policy. All the rest of the jobs residing in the system remain in the queue. Once a job receives the full amount of service it required, it departs the system. The epoch of departure is denoted d_i.

The service policy can rely on any of the jobs properties, including (but not limited to) the order of arrival, the service requirement, the time already spent in the system, the amount of service already given to the job, etc.

We limit our discussion to the following classes of service policies:

Φ_1 – The class of work conserving, non-idling service policies.

Φ_2 – The class of work conserving, non-idling, non-preemptive service polices. $\Phi_2 \subset \Phi_1$.

Work conserving requires that jobs receive exactly the amount of service they require, no less, no more.

Non-idling requires that while there are jobs in the system, the server will grant service at full service rate.

Non-preemptive requires that once the server started serving a job, it will not stop doing so until the job's service requirement is fulfilled. It also requires that at most one job be served at any epoch.

We use bar ($\overline{X}$) to denote expected value and hat ($\widehat{X}$) do denote variance.

3 Locality of Reference and Comparison Set

We now address the two difficulties presented in Section 1, focusing on the first example. It seems 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 demonstrated in Section 1, 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 unfairness among jobs. Intuitively speaking, the examples demonstrate 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". In other words, a job should be compared to any other job whose service scheduling can improve or worsen its treatment. Naturally, such an improvement or worsening is done by shifting resources between the two jobs. We now establish when such a resource shift is possible.

For a single-server work-conserving system (with no idling) we claim that the performance metrics 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.

This is established in the following theorem:

Theorem 3.1 (Locality of Reference)**.** *For any service policy in Φ_1, consider two arbitrary jobs J_a and J_b. The server can shift resources between them if and only if they reside in the same busy period.*

This theorem may seem intuitive, but requires detailed treatment. We thus provide the exact definition of resource shifting and the proof of this theorem in Appendix A (the theorem is given in Theorem A.1). In broad terms, we define resource shifting as the existence of an alternative assignment of service processing times in which a period exists for which service previously given to J_a is now given to J_b and vice versa, and the rest of the jobs are not affected.

This theorem suggests that the proper comparison set consists of the jobs of a busy period rather than the whole population.

4 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 unfairness evaluation it is desired to have the com-

parison set of the variance consisting of the jobs of a busy period, we now define two desired performance metric properties, designed to address the two difficulties raised in Section 1.

Assume that X is a random variable denoting a performance metric of an individual job (e.g., waiting time or sojourn time) when the system is at steady state. For an arbitrary job, let Y be a random variable denoting the value of the performance metric X averaged over the jobs participating in its busy period, when the system is under steady state.

Property 1 (Locality of Measurement)**.** *A performance metric is said to be* Locally Measured *if for every service policy in Φ_1, and for every sample path, Y is constant (not necessarily the same constant for all service policies).*

To understand this property, note that for many performance metrics X the corresponding Y random variable is not constant. For example, if X is the waiting time of a job then the average waiting time of the busy period is not constant and can vary from one busy period to another (e.g., consider FCFS applied on the busy periods in the example given in the introduction).

Intuitively speaking, if a performance 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). This provides a remedy to the difficulty demonstrated in the second example.

To address the second property we first define the concepts of *inter-variance* and *intra-variance*. The intra-variance of X , $\widehat{X}_{intra}$, is the second moment of X around the average of X *at the same busy period*, namely around Y. In formal terms, let J is a job and X is a random variable of a performance metric of the job (say, waiting time). Let b be an instant of a busy period determined by the number of jobs, n_b, their relative arrival times, $a_1, ..., a_{n_b}$, and their relative service requirements, $s_1, ..., s_{n_b}$.[1]. Let $f_B(b)$ be density distribution function of the busy period.

For a specific busy period instant b, the performance metric X over its jobs is given by specific values $x_1, ..., x_{n_b}$. Let $Y_{|b}$ denote the value of Y conditioned on $J \in b$. This is given by $Y_{|b} = \frac{1}{n_b} \sum_{i=1}^{n_b} x_i$ (this can also be written as $Y_{|b} = E[X|J \in b]$). Then the intra-variance is computed by first computing $E[(X-Y)^2|J \in b] = \frac{1}{n\,b} \sum_{i=1}^{n_b} (x_i - \frac{1}{n\,b} \sum_{i=1}^{n_b} x_i)^2$ (denoted $\widehat{X}^b_{intra}$) and then unconditioning on $J \in b$.[2] This is in contrast to the regular variance $\widehat{X}$, which we denote *global variance*, in which the second moment is computed around the expected value of X over the whole job population, i.e. $E[(X - \overline{X})^2]$.

The inter-variance $\widehat{X}_{inter}$ is defined as the second moment of Y around the expected value of X. In formal terms it is computed by taking $E[(Y - \overline{X})^2|J \in b] = (Y_{|b} - \overline{X})^2$ (which we denote $\widehat{X}^b_{inter}$) and then unconditioning on $J \in b$.

Intuitively speaking, the intra-variance measures the dispersion of specific measures from their average values in each busy period, while the inter-variance measures the dispersion of

[1]Note that by using the indices $1, ..., n_b$ to denote the indices of the jobs in the busy period we do not lose generality

[2]Note that unconditioning on $J \in b$ is done by $\frac{\int \widehat{X}^b_{intra} n_b f_B(b) db}{\int n_b f_B(b) db}$

the busy period average values.

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

Theorem 4.1. *$\widehat{X} = \widehat{X}_{intra} + \widehat{X}_{inter}$, or, in words, the global variance equals the sum of intra-variance and inter-variance.*

Proof. We start with conditioning $\widehat{X}$ on $J \in b$, which we denote $\widehat{X}^b$:

$$\begin{aligned}\widehat{X}^b =& E[(X-\overline{X})^2|J\in b] = E[\big((X-Y)+(Y-\overline{X})\big)^2|J\in b]\\ =& E[(X-Y)^2+(Y-\overline{X})^2+2(X-Y)(Y-\overline{X})|J\in b]\\ =& E[(X-Y)^2|J\in b]+E[(Y-\overline{X})^2|J\in b]+E[2(X-Y)(Y-\overline{X})|J\in b]\end{aligned}.$$

We now focus on the third term in the last line. Note that conditioned on $J \in b$, the second multiplicand is constant and therefore

$$E[2(X-Y)(Y-\overline{X})|J\in b] = 2(Y_{|b}-\overline{X})E[X-Y|J\in b] = 2(Y_{|b}-\overline{X})\cdot 0 = 0,$$

and thus

$$\widehat{X}^b = E[(X-Y)^2|J\in b]+E[(Y-\overline{X})^2|J\in b] = \widehat{X}^b_{intra}+\widehat{X}^b_{inter}.$$

Unconditioning on $J \in b$ we get $\widehat{X} = \widehat{X}_{intra} + \widehat{X}_{inter}$. □

We can now define the second property:

Property 2 (Locality of Variance)**.** *A performance metric is said to have its* variance local *if its global variance equals its intra-variance for every service policy in Φ_1, and for every sample path, i.e. $\widehat{X} = \widehat{X}_{intra}$.*

Such a performance metric does not suffer from the difficulty presented in the first example, as there are no differences between busy periods, and thus the only differences measured are those within busy periods.

Properties 1 and 2 are strongly related:

Theorem 4.2. *A performance metric is locally measured if and only if its variance is local.*

Proof. If a performance metric is locally measured then Y is constant, i.e. X averaged over the jobs participating in a busy period is the same for all busy periods. Therefore Y is constant, and clearly, $Y_{|b} = \bar{X}$, so the inter-variance is zero. From Theorem 4.1 it follows that the variance of the performance metric is local.

On the other direction, if a performance metric has its variance local then from the definition of Property 2 and from Theorem 4.1 it follows that it has zero inter-variance. From the definition of inter-variance, this means that $(Y_{|b}-\overline{X})^2 = 0 \Rightarrow Y_{|b} = \overline{X}$, and thus $Y_{|b}$ is the same constant for all busy periods, and clearly Y is constant, satisfying the requirement for being locally measured. □

5 Locally Measured Metrics

We now investigate for which performance metrics the properties defined above hold.

We consider a class of performance metrics ξ which is relatively large. Let $N(t)$ be number of jobs in the system at time t. Each performance metric $X \in \xi$ is identified by two functions, the warranted service function $f(N)$ and the utility function $g(S)$. The performance metric is

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

Intuitively, X_i is the net amount of utility job J_i receives, which is the total utility $g(s_i)$ minus the utility the job is warranted due to its stay in the system, which is momentarily determined by $f(N(t))$.

Note that class ξ is quite wide and includes a variety of performance metrics. For example, the waiting time and the sojurn time metrics both belong to ξ.

Specifically, we now examine the *discrimination function* proposed in [4]. This performance metric aims at evaluating the deviation of the service a job receives from what it "deserves" to receive from an equal sharing of resources point of view. For job J_i the performance metric is $D_i = s_i - \int_{a_i}^{d_i} \frac{dt}{N(t)}$. The utility function in this case is the unit function $g(S) = 1$, i.e. the utility is one unit per unit of service time, and the warranted service function is a fair share, i.e. $f(N) = 1/N$. In [4] it was proposed to use the variance of the random variable D (under steady state) as a metric of system unfairness. It is easy to see that $D \in \xi$.

We further define a class of performance metrics $\mathcal{D} \subset \xi$ where $f(N) = \alpha/N$ and $g(S) = \alpha S + \beta$, where α and β are constants and $\alpha, \beta \geq 0$. The performance metric is

$$X_i = \beta + \alpha s_i - \int_{a_i}^{d_i} \frac{\alpha dt}{N(t)}. \tag{1}$$

Intuitively speaking, this generalization serves for a case where the job gets a certain fixed benefit β from being served, and an additional constant benefit of α per unit of service time. Choosing $\alpha = 1$ and $\beta = 0$ yields the discrimination function D defined in (1). Note that since we are dealing with work conserving systems $s_i = \int_{a_i}^{d_i} s_i(t)dt$ where $s_i(t)$ is the rate of service given to job i at epoch t. We can therefore write the performance metric in the following way

$$X_i = \int_{a_i}^{d_i} \alpha \left(s_i(t) - \frac{1}{N(t)} \right) dt + \beta. \tag{2}$$

We now move on to show the importance of the class $\mathcal{D}$

Theorem 5.1. *For every performance metric $X \in \mathcal{D}$ both Property 1 and Property 2 hold.*

Proof. Let X be a performance metric such that $X \in \mathcal{D}$. Let Y be a random variable denoting the value of the performance metric X averaged over the jobs participating in its busy period, when the system is under steady state.

Consider an arbitrary busy period b with n_b jobs. Assume, without loss of generality that the job indices are $1, 2, ..., n_b$ and let d_{last} denote the last departure epoch of the busy period. The expected value of X in this busy period, $\overline{X}^b$, is

$$\overline{X}^b = \frac{1}{n_b}\sum_{i=1}^{b_b}\left(\int_{a_i}^{d_i}\alpha\left(s_i(t)-\frac{1}{N(t)}\right)dt+\beta\right) = \beta+\frac{\alpha}{n_b}\int_{a_1}^{d_{last}}\sum_{i|t\in(a_i,d_i)}\left(s_i(t)-\frac{1}{N(t)}\right)dt$$

$$= \beta+\frac{\alpha}{n_b}\int_{a_1}^{d_{last}}\left(\sum_{i|t\in(a_i,d_i)}s_i(t)-\sum_{i|t\in(a_i,d_i)}\frac{1}{N(t)}\right)dt, \quad (3)$$

where we evaluate the performance metric using (2).

Consider the first sum in the integral. Note that the system is non-idling, and thus the total amount of service given in any epoch is unity. Considering the second sum, we note that the condition $i|t\in(a_i,d_i)$ holds for exactly $N(t)$ jobs and thus $\sum_{i|t\in(a_i,d_i)}\frac{1}{N(t)} = N(t)\frac{1}{N(t)} = 1$.

Substituting the above into (3) we get

$$\beta+\frac{\alpha}{n_b}\int_{a_1}^{d_{last}}(1-1)dt = \beta.$$

Since β is constant, Y is a constant and Property 1 holds, and from Theorem 4.2 so does Property 2. □

We now move on to show the uniqueness of $\mathcal{D}$, that is:

Theorem 5.2. *Let $X\in\xi$. If either Property 1 or Property 2 holds for X, then $X\in\mathcal{D}$.*

Proof. We show that if Property 1 holds for $X\in\xi$, then $X\in\mathcal{D}$. The parallel claim for Property 2 will follow from Theorem 4.2.

Note that for Property 1 to hold for X, then for *every service policy* $\phi\in\Phi_1$ the measure Y of an arbitrary job must be constant (though, of course, this constant can be different for different service policies). For a performance metric $X\in\xi$ we will assume that Y is a constant for the processor sharing policy (PS) and show that $X\in\mathcal{D}$.

Recall that Y is a measure attributed to an arbitrary job J and is equal to the average of X over the jobs participating with J in the same busy period. Let Z be the average of X taken over the jobs of an *arbitrary busy period*. It immediately follows that if Y is a constant, then Z is a constant as well. We will now construct a busy period, and as X satisfies Property 1, the average of X taken over the jobs in this busy period must be a constant.

Consider a busy period, say j, that starts with the simultaneous arrival of N^j identical jobs, each with the same service requirement s^j. According to PS all the jobs are served

concurrently for duration $N^j s^j$. As all the jobs have identical service requirements and all reside at the system exactly at the same epochs, the performance measures for all these jobs are equal, that is $X_i = X_k$ for every i, k and are exactly

$$g(s^j) - \int_{a_i}^{d_i} f(N^j(t))dt = g(s^j) - N^j s^j f(N^j),$$

since $N^j(t) = N^j$ is constant, and $d_i - a_i = N^j s^j$. As all individual measures are equal this also equals $\overline{X}^j$

As X satisfies Property 1 this value is constant for every busy period, say c, and thus

$$g(s^j) - N^j s^j f(N^j) = c \Rightarrow N^j f(N^j) = \frac{g(s^j) - c}{s^j}. \tag{4}$$

Note that no restriction was set on N^j and s^j, and therefore this must be true for every value of N^j and s^j, and *with the same constant* c. Therefore, the left hand side of the equation cannot depend on N^j and the right hand side cannot depend on s^j. This can only be satisfied if $g(s^j) = c + ds^j$ and $f(N^j) = d/N^j$, where d is a constant. We can now replace d by α, c by β, s^j by S and N^j by N to arrive at the exact definition of $\mathcal{D}$, i.e. $f(N) = \alpha/N$ and $g(S) = \alpha S + \beta$. □

Note that our proof is correct because of the definition of Property 1, namely that a constant should be provided for *every* service policy in Φ_1. Consider now a weaker definition of Property 1 in which it is required that a constant be provided for *some* service policy in Φ_2 (recall that $\Phi_2 \subset \Phi_1$) instead of *every* service policy in Φ_1:

Property 1a (Weak Locality of Measurement)**.** *A performance metric is said to be* Weakly Locally Measured *if there exists a service policy $\phi \in \Phi_2$ such that for every sample path, Y is constant.*

Property 2 can be similarity defined (say, Property 2a), and the two properties can be shown to be identical through a theorem similar to Theorem 4.2. For these definitions, a stronger claim can be proved:

Theorem 5.3. *Let $X \in \xi$. If either Property 1a or Property 2a hold for X, then $X \in \mathcal{D}$.*

Intuitively speaking, a performance metric satisfying Property 1 can be used to measure and compare every service policy in Φ_1. According to Theorem 5.2, within ξ, only performance metrics belonging to $\mathcal{D}$ satisfy this requirement. However, some performance metric may exist that can be used to measure and compare *only a subset* of Φ_1, and not the entire class. Theorem 5.3 shows that if this subset includes even a single policy in Φ_2 (and recall that the most common policy FCFS, belongs to Φ_2), then again, this performance metric must be in $\mathcal{D}$.

As the proof is lengthy it is brought in Appendix B.

6 Explicit Evaluation of the Intra-Variance

This paper started explaining why using the waiting time (sojourn time) variance as a measure of unfairness can lead to difficulties, and that only the variability within a busy period (later defined as the *intra-variance*) matters for the purpose of fairness. Our discussion led to the uniqueness of $\mathcal{D}$ within ξ, in having a variance equal to the intra-variance, thus avoiding this problem. A second approach one can use for devising a fairness metric is to choose a performance metric, say the sojourn time, and *explicitly* evaluate its intra-variance as an unfairness metric.

To demonstrate how this solves the difficulty, let us readdress the example given in the introduction. Recall that arrivals occur in bulks consisting of either 2 jobs in a bulk (type A) or 4 jobs in a bulk (type B), with equal probability for either bulk, and the server processes the jobs in a Processor Sharing mode. The sojourn time of all jobs are therefore 2 units and 4 units for type A and type B, respectively. To calculate the inter-variance, we evaluate $E[(X-Y)^2|J \in b]$ for every $J \in b$. In all cases this equals zero as $X = Y_{|b}$. Unconditioning on $J \in b$ yields an intra-variance of zero, or no unfairness, which is proper for this system.

While this method seems viable it has two major drawbacks. First, for the customer, this method does not address the issue of locality of measurement at all, i.e. it does not provide a customer with a scale of reference to determine whether its specific job is positively or negatively discriminated, and how much. To address this issue one needs to know the average measure in each busy period in an on-line manner, while the information is only available at the end of busy period.

Second, for the analyst, while the intra-variance can be easily obtained through simulation, it makes analytical results hard to obtain. To compute the intra-variance through simulation, one needs only to compute for every job its performance metric (say waiting time) and for every busy period its average performance metric (average waiting time), derive the difference squared, and average over all jobs. However, if one wishes to carry out an analysis of this measure, one needs to compute the following expression:

$$\begin{aligned} E[(X-Y)^2|J \in b] &= E[X^2 + Y^2 - 2XY|J \in b] \\ &= E[X^2|J \in b] + E[Y^2|J \in b] - 2E[XY|J \in b] \\ &= E[X^2|J \in b] + (E[Y|J \in b])^2 - 2E[Y|J \in b]E[X|J \in b]' \\ &= E[X^2|J \in b] - (E[Y|J \in b])^2 \end{aligned}$$

(where the third line results from the fact that, conditioned on $J \in b$, Y is a constant) and then unconditioning on $J \in b$. While the first term simply yields the second moment of X, the second term does not yield to analysis in most practical cases of randomly sampled arrival times and service requirements. We leave this as a challenging open problem. In comparison, RAQFM is analyzable for at least $M/PH/m$ (see [2]).

7 Numerical Results

In this section we present numerical results from a simple scenario and from simulating some simple stochastic systems.

7.1 Numerical Example

We start with a simple numerical example that shows that using different metrics can lead to contradicting results. We compare the (global) variance of the sojourn time to the (global) variance of the discrimination and show that the results are contradicting. We then compare the (global) variance of the sojourn time to its intra-variance and show again that the results contradict.

Consider a system where there are two types of customers A and B, with service requirements $s_A = a$ and $s_B = b$, respectively. The arrival rate is very low ($\to 0$) and each arrival consists of two customers of the same type, with probability of 0.5 to any of the two types. We examine two service policies, PS and FCFS (in this specific case all non-preemptive policies behave the same).

We start with comparing the variance of the discrimination to that of the sojourn time(Table 1).

		Discrimination	Sojourn Time
Global Variance	FCFS	$\frac{1}{8}(a^2+b^2)$	$\frac{1}{16}(11a^2-18ab+11b^2)$
	PS	0	$(a-b)^2$
Intra-Variance	FCFS	$\frac{1}{8}(a^2+b^2)$	$\frac{1}{8}(a^2+b^2)$
	PS	0	0
Inter-Variance	FCFS	0	$\frac{9}{16}(a-b)^2$
	PS	0	$(a-b)^2$

Table 1: Variances of Discrimination Vs. Variance of Sojourn Time

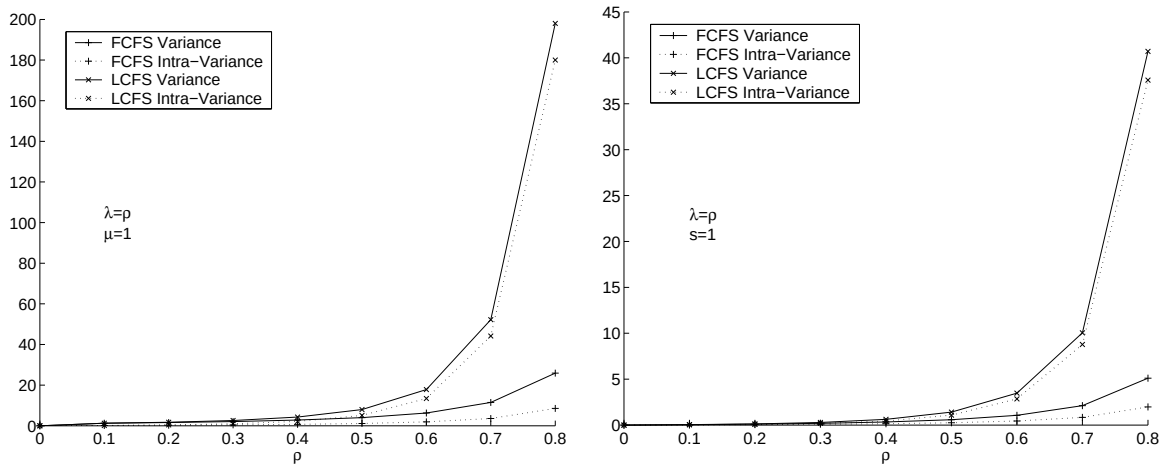
Using the variance of discrimination as a fairness measure, clearly, $\frac{1}{8}(a^2+b^2) > 0$, thus PS is more fair than FCFS. Using the variance of the sojourn time we see that $\frac{1}{16}(11a^2 - 18ab - 11b^2)$ is not always larger than $(a-b)^2$. In fact, for every choice of a, if we choose $b > \frac{1}{5}(7+2\sqrt{6})a \cong 2.38a$ it yields as a result that FCFS is more fair than PS. Specifically, if we choose $a = 1, b = 3$, the variance of the sojourn time is 3.5 for FCFS and 4 for PS. The variance of discrimination, in contrast, is 1.25 for FCFS and 0 for PS and $1.25 > 0$. As for the intra-variance of the sojourn time we see that $(a-b)^2 > 0$, meaning that according to the intra-variance of the sojourn time PS is more fair than FCFS, in agreement with the variance of discrimination.

As expected from Theorem 4.1, the (global) variance equals the sum of the intra-variance and the inter-variance, $\frac{1}{8}(a^2+b^2) + \frac{9}{16}(a-b)^2 = \frac{1}{16}(11a^2 - 18ab - 11b^2)$.

In conclusion, we see that using the variance of the sojourn time as a fairness measure leads to FCFS being more fair than PS. This is caused by the sojourn time not having a local variance. Using either the variance of discrimination, or explicitly using the intra-variance of the sojourn time leads to the opposite, and perhaps more appropriate, result.

7.2 Simulation Results

Figure 1 depitcs simulation results comparing the (global) variance of the sojourn time to its intra-variance, in the following settings. Figure 1(a) and Figure 1(b) present results for exponential service requirement distribution and constant (unit) service requirements, respectively. In each case we compare the FCFS policy to the LCFS one. Each of the plotted points is the result of simulation of at least 10^6 jobs.



(a) Exponential Service Requirement Distribution (b) Constatnt Service Requirements

Figure 1: Variance Vs. Intra-Variance of Sojourn Time

The results demonstrate that indeed, the variance and the intra-variance differ significantly. This is more prominent for FCFS, and can reach relative differences of 100 percents or more, but is evident in both service policies and for both service requirement distributions.

8 Concluding Remarks

We showed in this work that there are two main difficulties with using the variance of sojourn time as an unfairness metric. The first is that it involves comparisons between jobs which are not related, and the second is that performance metrics lack a global point of reference. We then showed that RAQFM (the global variance of the discrimination D) is not subject to these difficulties, and is unique in this property within a large class of performance metrics. Thus, one may conclude that using the (global) variance of D is appropriate for quantifying

queue unfairness, while using the (global) variance of other performance metric (such as the variance of sojourn time) may sometimes lead to inaccuracies.

We noted that the (analytic) derivation of the global variance of a performance metric is typically much simpler than that of the local variance.

An interesting open problem is whether the class of performance metrics considered, ξ, can be made more general, making the statements in this work stronger. We suspect it can be done.

We would also like to point out that it might be possible to construct specific performance metrics that will be locally measured for 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 restrictions will make those performance metrics impractical.

References

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

[2] E. Brosh, H. Levy, and B. Avi-Itzhak. The effect of service time variability on job scheduling fairness. Technical Report RRR-24-2005, RUTCOR, Rutgers University, July 2005.

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

[4] 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).

[5] 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.

A Proof of the Locality of Reference Theorem

In this appendix we provide the Locality of Reference theorem and its proof in details.

Definition A.1. *An Arrival and Service Pattern (or "a pattern" in short) is a series of arrival times and service requirements* $\{(a_i, s_i)\}_{i=1}^{\infty}$ *corresponding do a series of jobs* $J_1, J_2, \ldots$.

Definition A.2. *A Resource Allocation Schedule (RAS) is a function* $\varphi\colon [0,\infty) \to \mathbb{N} \cup \{0\}$ *that for each epoch* $t \in [0,\infty)$ *returns the index* ($\varphi(t)$) *of the job served at that epoch or* 0 *if the server is idle.*

Remark A.1. (Processor Sharing Policies): *Note that while the definition of RAS limits our model to serving one job at each epoch, it can still be used to model any processor sharing type system, by using infinitesimally small intervals.*

For a given pattern we will say that RAS φ belongs to Φ_1 if it satisfies the following requirements:

1. Arrival Validity: $\forall t, \varphi(t) = k \Rightarrow t \geq a_k$ i.e. no job is given service before its arrival.
2. Work Conserving: $\forall k, \int\limits_{\varphi(t)=k} dt \leq s_k$ i.e. no job is served more than its service requirement
3. Non-Idling: $\forall t, \varphi(t) = 0 \Rightarrow \forall k, a_k < t\colon \int\limits_{\substack{a_k \\ \varphi(u)=k}}^{t} du = s_k$ i.e., the server does not idle at t if there is a job in the system that still requires service.

Note that requirement 3 implies that for each J_k there exists $\epsilon > 0$ such that

$$\forall t \in (d_k - \epsilon, d_k)\colon \varphi(t) = k. \tag{5}$$

Definition A.3. *Consider RAS φ_a and RAS φ_b. We say that φ_a and φ_b are 1-swappable for a pair of jobs $\langle J_i, J_j \rangle$ if:*

$$\begin{aligned} \exists t_1, t_2, \varepsilon\colon \quad & (\forall t \in (t_1, t_1 + \epsilon)\colon \varphi_a(t) = i, \varphi_b(t) = j), \\ & (\forall t \in (t_2, t_2 + \epsilon)\colon \varphi_a(t) = j, \varphi_b(t) = i), \\ & (\forall t \notin (t_2, t_2 + \epsilon) \cup (t_1, t_1 + \epsilon) : \varphi_a(t) = \varphi_b(t)) \end{aligned}$$

In other words, the RASs are identical except for a shift of resources between $\langle J_i, J_j \rangle$ at the segments $(t_1, t_1 + \epsilon)$ and $(t_2, t_2 + \epsilon)$.
We recursively define φ_a and φ_b to be n-swappable for $\langle J_i, J_j \rangle$ if there exists a RAS φ_c and a job J_k such that φ_a and φ_c are $(n-1)$-swappable for $\langle J_i, J_k \rangle$ and φ_c and φ_b are 1-swappable for $\langle J_k, J_j \rangle$.
We say that φ_a and φ_b are resource-swappable for $\langle J_i, J_j \rangle$ if there exists a value $n \in \mathbb{N}$ such that they are n-swappable for $\langle J_i, J_j \rangle$.

Lemma A.1. *For every job J_i that is not the first job in a busy period, there exists another job J_j such that $a_j < a_i, d_j > a_i$.*

Proof. Assume there is no such job. Then every job arriving before J_i also leaves before J_i arrives, and that makes J_i the first in a busy period, in contradiction to the assumption. □

Theorem A.1 (Locality of Reference)**.** *Let $\{(a_i, s_i)\}_{i=1}^{\infty}$ be an arbitrary pattern, let φ_a be an arbitrary RAS that belongs to Φ_1 for that pattern, and let $\langle J_i, J_j \rangle$ be an arbitrary pair of jobs. Then, J_i and J_j are in the same busy period if and only if there exists a RAS φ_b that belongs to Φ_1 for that pattern, and φ_a and φ_b are resource-swappable for $\langle J_i, J_j \rangle$.*

Proof. i) Assume that J_i and J_j are in the same busy period and prove the existence of RAS φ_b. We denote the departure epoch of J_k under φ_a and φ_b as d_k^a and d_k^b respectively.

Assume $d_i^a < d_j^a$. We will prove the claim by induction over $a_j - d_i^a$.

Induction base: $a_j - d_i^a < 0$. Then according to (5) there exists ϵ_1 such that $\forall t \in (d_i^a - \epsilon_1, d_i^a)$: $\varphi(t) = i$ and there exits ϵ_2 such that $\forall t \in (d_j^a - \epsilon_2, d_j^a)$: $\varphi(t) = j$. We define $\epsilon = \min\{\epsilon_1, \epsilon_2\}$ and construct φ_b as follows:

$$\varphi_b(t) = \begin{cases} j & t \in (d_i^a - \epsilon, d_i^a) \\ i & t \in (d_j^a - \epsilon, d_j^a) \\ \varphi_a(t) & otherwise \end{cases}.$$

It is easy to see that φ_a and φ_b are 1-swappable for $\langle J_i, J_j \rangle$ at the segments $(d_i^a - \epsilon, d_i^a)$ and $(d_j^a - \epsilon, d_j^a)$.

Induction step: we assume that if $a_j - d_i^a < l$ there exists φ_b such that φ_a and φ_b are resource-swappable for $\langle J_i, J_j \rangle$ and show that if $a_j - d_i^a = l$ there also exists such a RAS.

Assume $a_j - d_i^a = l$. We choose J_m such that $a_m < a_j, d_m > a_j$, which must exist according to Lemma A.1. As $a_m < a_j$ we have $a_m - d_i^a < l$ and according to the induction assumption there exists a RAS, say φ_c, such that φ_a and φ_c are resource-swappable for the pair $\langle J_m, J_j \rangle$.

We now show that there exists a RAS φ_b such that φ_c and φ_b are 1-swappable for $\langle J_m, J_j \rangle$, and conclude that φ_a and φ_b are resource-swappable.

Let τ be the first epoch in which J_j is served according to φ_c, and assume it is served for a period of ϵ_1. According to (5) there exits ϵ_2 such that $\forall t \in (d_m^c - \epsilon_2, d_m^c)$: $\varphi(t) = m$. We define $\epsilon = \min\{\epsilon_1, \epsilon_2\}$ and construct as follows:

$$\varphi_b(t) = \begin{cases} j & t \in (d_m^c - \epsilon, d_m^c) \\ m & t \in (\tau, \tau + \epsilon) \\ \varphi_c(t) & otherwise \end{cases}.$$

It is easy to see that φ_b and φ_c are 1-swappable for $\langle J_m, J_j \rangle$ at the segments $(d_m^c - \epsilon, d_m^c)$ and $(\tau, \tau + \epsilon)$.

The proof for the case $d_i^a > d_j^a$ is similar.

Remark A.2 (Validity of the Induction)**.** *While the induction is defined on a continuous space (real numbers) it does not require infinite number of steps. This can be shown to be true as the choice of J_m requires $a_m < a_j$. This can be done at most a number of times equal to the number of customers in the system. Furthermore, let ϵ_a be the minimum inter arrival interval length in the busy period. Then a sufficient assumption for the induction step is that if $a_j - d_i^a \leq l - \epsilon_a$ there exists φ_b such that φ_a and φ_b are resource-swappable for $\langle J_i, J_j \rangle$, thus each step reduces l by at least ϵ_a.*

This concludes the proof of the first direction.

ii) To prove the other direction of the theorem one has to show that if J_i and J_j are not in the same busy period then there exists no φ_b such that φ_a and φ_b are resource-swappable

for $\langle J_i, J_j \rangle$. This results from the fact that the server cannot do resource swapping across different busy periods since all the jobs of one busy period leave the system before any jobs of the other busy period arrive. This concludes the proof. □

B Proof of the Uniqueness of $\mathcal{D}$ under Property 1a

Theorem 5.3. *Let $X \in \xi$. If either Property 1a or Property 2a hold for X, then $X \in \mathcal{D}$.*

Proof. Consider an arbitrary service policy $\phi \in \Phi_2$. Assume that Property 1a holds for a performance metric $X \in \xi$ for this arbitrary service policy, and for every sample path. We show that $X \in \mathcal{D}$.

Consider a specific type of busy periods, which we will name the *Heavy First Job* scenario. Consider a busy period, say j, consisting of N^j jobs. The first job to arrive at the system has a "heavy" service requirement of $N^j s^j$ units of time. Following this job are $N^j - 2$ jobs, each with a service requirement of s^j units of time. These jobs arrive every s^j units of time, starting s^j units of time after the beginning of the busy period. In addition, one job arrives at the system ϵ units of time before the last job completes service with a service requirement of ϵ, where $\epsilon \to 0$.

As this is a non-preemptive service policy, the first job will be served until finished. Following this, the other jobs will be served in some arbitrary order defined by the service policy. However, as those jobs are identical in their service requests, the order of service does not influence $N(t)$, which is depicted in Figure 2. We now evaluate $\overline{X}^j$. It is useful at this

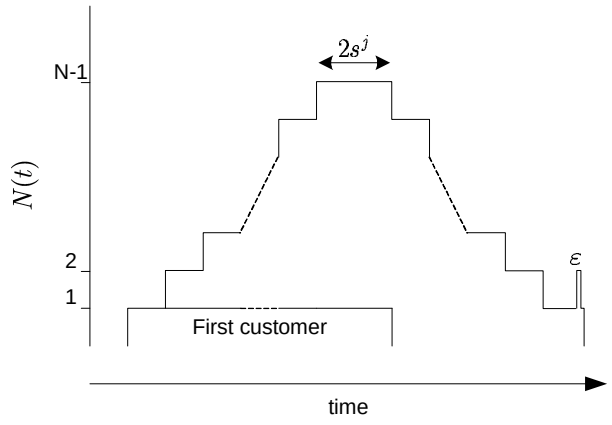


Figure 2: A non-preemptive service policy with a "Heavy First Job" scenario

point to introduce an alternative way of calculating $\overline{X}^j$. Let d_{last} denotes the last departure

epoch of the busy period, then

$$\begin{aligned}\overline{X}^j &= \tfrac{1}{N^j}\sum_{i=1}^{N_j}\left(g(S_i) - \int_{a_i}^{d_i} f(N(t))dt\right) \\ &= \tfrac{1}{N^j}\left(\sum_{i=1}^{N_j} g(S_i) - \int_{a_1}^{d_{last}}\left(\sum_{i|a_i\le t\le d_i} f(N(t))\right)dt\right) . \\ &= \tfrac{1}{N^j}\left(\sum_{i=1}^{N_j} g(S_i) - \int_{a_1}^{d_{last}} N(t)f(N(t))dt\right),\end{aligned} \tag{6}$$

Note that the the warranted service is derived by integrating $N(t)f(N(t))$ over the entire busy period.

The value of (6) for the specific arrival and service pattern we consider is:

$$\overline{X}^j = \frac{1}{N^j}\left((N^j-2)g(s^j) + g(N^j s^j) + g(\epsilon) - 2s^j\sum_{i=1}^{N^j-1} if(i) - \epsilon f(2)\right),$$

which after substituting $\epsilon \to 0$ yields

$$\overline{X}^j = \frac{1}{N^j}\left((N^j-2)g(s^j) + g(N^j s^j) + g^* - 2s^j\sum_{i=1}^{N^j-1} if(i)\right),$$

where $g^* = \lim_{S\to 0} g(S)$, which is constant for every function $g(S)$.

Now consider a similar busy period, say k, except that on this busy period instead of the last job (the one with the service requirement of ϵ units of time) a regular job with a service requirement of s^j units of time arrives $(N^j-1)s^j$ units of time after the beginning of the busy period. Figure 3 depicts $N(t)$, where the gray squares form the difference between the busy periods.

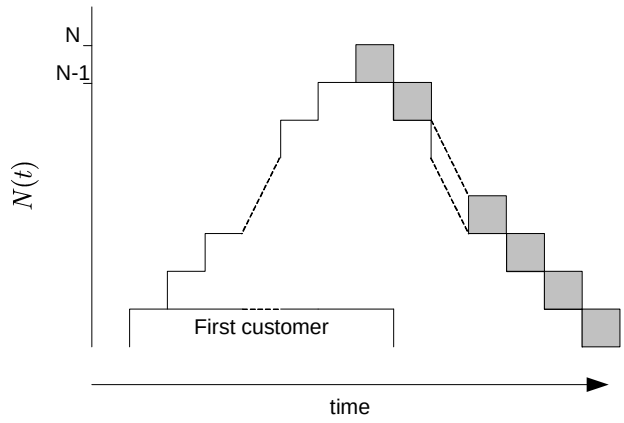


Figure 3: The "Heavy First Job" scenario 2

For this busy period

$$\overline{X}^k = \frac{1}{N^j}\left((N^j-1)g(s^j) + g(N^j s^j) - 2s^j \sum_{i=1}^{N^j-1} i f(i) - s^j N^j f(N^j)\right).$$

As Property 1a holds for X, $\overline{X}^j = \overline{X}^k$ leading to

$$\overline{X}^k - \overline{X}^j = \frac{1}{N^j}\left(g(s^j) - g^* - s^j N^j f(N^j)\right) = 0 \Rightarrow N^j f(N^j) = \frac{g(s^j) - g^*}{s^j}.$$

This bears a striking resemblance to (4) as g^* is constant for $g(S)$. We follow the same reasoning used in the proof of Theorem 5.2 to complete the proof. □