Adv. Appl. Prob. **22,** 257–259 (1990)
Printed in N. Ireland

THE EQUALITY OF THE VIRTUAL DELAY AND ATTAINED WAITING TIME DISTRIBUTIONS

HIROTAKA SAKASEGAWA,* *University of Tsukuba*
RONALD W. WOLFF,** *University of California, Berkeley*

Abstract

It has recently been shown that for the $G/G/1$ queue, virtual delay and attained waiting time have the same stationary distribution. We present a sample-path derivation of this result.

G/G/1 QUEUE; FIFO; FCFS; WORK IN SYSTEM

Sengupta (1989) has recently shown that for the first-in–first-out (FIFO) $G/G/1$ queue, virtual delay and attained waiting time (defined below) have the same stationary distribution. His proof is based on relationships between stationary quantities in Miyazawa (1979), (1983).

In this paper, we present a simple and direct sample-path proof of this result. In fact, our result is more general because it depends only on the existence and relative magnitude of two limits. In what follows, we fix a point ω in the sample space, and treat all quantities as numbers.

Let customer C_n arrive at epoch t_n and have service time S_n, $n \geqq 1$, where $0 \leq t_1 \leqq t_2 \cdots$. Customers are served at a single-server FIFO queue. Let *virtual delay* (*work in system*) $V(t)$ be the sum of the remaining service times of all customers in system at epoch $t \geqq 0$, and V_n be the *work found* by C_n, which is the sum of the remaining service times of all customers C_j, $j < n$, at epoch t_n. If customers arrive one at a time, $V_n = V(t_n^-)$. For arbitrary fixed $V(0)$, $V_1 = (V(0) - t_1)^+$.

For every fixed $V(0)$, assume that the system empties infinitely often. A sufficient condition for this is that the following limits exist:

$$\lim_{n\to\infty} t_n/n = 1/\lambda \quad \text{and} \quad \lim_{n\to\infty} \sum_{j=1}^{n} S_j/n = 1/\mu, \quad \text{where} \quad 0 < \lambda < \mu < \infty. \tag{1}$$

Now C_n enters service at epoch $\tau_n = t_n + V_n$, and we define the *attained waiting time* $W_a(t)$ of the customer in service at epoch t as

$$W_a(t) = t - t_n = V_n + t - \tau_n \quad \text{for} \quad t \in [\tau_n, \tau_n + S_n), \tag{2}$$

which is the work found by C_n (C_n's *delay in queue*) plus C_n's *attained service* at epoch t. Define $W_a(t) = 0$ if the system is empty epoch t.

Notice that the *area* under $\{W_a(t)\}$ while C_n is in service is a trapezoid with base S_n, left side V_n, and right side $V_n + S_n$. We find a corresponding area under $\{V(t)\}$. For this purpose, we *pretend* that the queue is operated under *preemptive LIFO* (PL), where the server is always serving the most recent arrival (the customer C_n with the largest subscript), among those in system. PL has no effect on the sample paths of $\{V(t)\}$. Under PL, consider the area

Received 10 August 1989.

* Postal address: Institute of Socio-Economic Planning, The University of Tsukuba, Tsukuba, Ibaraki 305, Japan.

** Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA.

Part of this research was done while R. W. Wolff visited the Department of Information Sciences, Science University of Tokyo, in 1989.

under $\{V(t)\}$ while C_n is in service. Because service on C_n may be interrupted, this area may have disjoint pieces. Put the pieces together! We again have a trapezoid with the same base S_n, but now the *right* side is V_n, and the *left* side is $V_n + S_n$. It follows immediately that while C_n is in service (under FIFO for W_a and PL for V), the amount of time for which $V(t) > x$ is equal to the amount of time for which $W_a(t) > x$. Adding up these amounts, we have the following result.

Theorem 1. For every busy period and every $x \geqq 0$, the amount of time for which $V(t) > x$ is equal to the amount of time for which $W_a(t) > x$.

To convert Theorem 1 into a statement about time averages, let $\{W'_a(t)\}$ be the process we get when the 'high' and 'low' sides of the trapezoids under $\{W_a(t)\}$ are reversed, and let $I_{vx}(t)$, $I_{wx}(t)$, and $I'_{wx}(t)$ be the indicators of the events $V(t) > x$, $W_a(t) > x$, and $W'_a(t) > x$, respectively; $t, x \geqq 0$.

We now compare the integrals $\int_0^t I_{\cdot}(u)\, du$. The processes V and W'_a have the same trapezoids, where the trapezoid of C_n begins earlier for V, and may end later, but only if work exceeds V_n in between. Thus the contribution of C_n to I_{vx} on $[0, t]$ either occurs earlier, or the work exceeds x throughout $[t_n, t]$. The trapezoids under W and W'_a are in the same locations, with the high sides under W'_a on the left. It follows immediately that

$$\int_0^t I_{vx}(u)\, du \geqq \int_0^t I'_{wx}(u)\, du \geqq \int_0^t I_{wx}(u)\, du. \tag{3}$$

Because the completed portions of the trapezoids under V in the interval $[0, t - W_a(t)]$ correspond to customers who arrived earlier than the (FIFO) customer in service at epoch t, we have the left-hand inequality in (4), which is combined with (3), for $t > 0$ and $x \geqq 0$,

$$\int_0^{t-W_a(t)} I_{vx}(u)\, du/t \leqq \int_0^t i_{wx}(u)\, du/t \leqq \int_0^t I_{vx}(u)\, du/t. \tag{4}$$

Now let $t \to \infty$ in (4). Both processes have the same time average, when the limits exist, provided that

$$W_a(t)/t \to 0 \quad \text{as} \quad t \to \infty. \tag{5}$$

We show that (1) implies (5) by first obtaining a result for the unstable system obtained by adding constant $c = 1/\lambda - 1/\mu + \varepsilon \geqq 0$ to all of the service times. Let V_{nc} be the work found by C_n with these altered service times. Let $X_n = S_n + c - (t_{n+1} - t_n)$, where $\lim_{n\to\infty} \sum_{j=1}^n X_j/n = \varepsilon$.

Now $\{V_{nc}\}$ is simply the standard delay-in-queue sequence for a single-server queue, where

$$V_{n+1,c} = (V_{nc} + X_n)^+, \quad \text{and} \quad Y_n = (V_{nc} + X_n)^-, \qquad n \geqq 1, \tag{6}$$

where Y_n is the idle time of the server between the completion of service on C_n and the commencement of service on C_{n+1}. From (6), we can write

$$V_{n+1,c} = V_1 + \sum_{j=1}^n x_j + \sum_{j=1}^n Y_j, \qquad n \geqq 1.$$

If c is large enough to make $\varepsilon > 0$, there is an m such that $\sum_{j=1}^n X_j > 0$ for all $n \geqq m$, which means that the queue remains busy from that point on, $Y_n = 0$ for all $n > m$, and we can write

$$V_{n+1,c} = V_1 + \sum_{j=1}^n X_j + \sum_{j=1}^m Y_j, \qquad n \geqq m. \tag{7}$$

Now divide (7) by n, and let $n \to \infty$. We have

$$\lim_{n\to\infty} V_{nc}/n = \varepsilon > 0, \tag{8}$$

where ε may be arbitrarily small.

https://doi.org/10.2307/1427611 Published online by Cambridge University Press

Now the V_{nc} are monotone non-decreasing in c for every n, and it follows immediately from (8) that when the limits in (1) hold under the slightly weaker conditions $0 < \lambda \leqq \mu < \infty$,

$$\lim_{n\to\infty} V_n/n = 0. \tag{9}$$

Remark. Our argument for (8) and (9) is similar to that in Loynes (1962). His results are in a stochastic setting, where $\{X_n\}$ is stationary and ergodic, but his approach is really sample-path in nature.

For $t > 0$, let $C_{n(t)}$ be the customer in service at epoch t under FIFO, and write

$$W_a(t)/t \leqq (V_{n(t)} + S_{n(t)})/t_{n(t)}, \tag{10}$$

where if the system is empty at epoch t, we set $t_{n(t)} = t$, and the other quantities in (10) equal to 0. Now $n(t) \to \infty$ as $t \to \infty$, and from (1), $t_{n(t)}/n(t) \to 1/\lambda$ as $t \to \infty$, and $S_n/n \to 0$ as $n \to \infty$. Thus, from (9) and (10), it is easy to see that (1) implies (5).

From (4) and that (1) implies (5), we have the following result.

Theorem 2. When (1) holds, the fraction of time that $\{V(t)\}$ exceeds x is equal to the fraction of time that $\{W_a(t)\}$ exceeds x, for every $x \geqq 0$, whenever these time averages exist as limits. If one time average exists, so does the other.

In a stochastic setting, where these time averages exist as constants w.p.1, and each process has a stationary distribution, each distribution must equal the correponding time average. Hence, in the stochastic setting in Sengupta (1989), Theorem 2 is equivalent to his result. Also note that in our analysis, we do not require either arrivals or departures to be one at a time.

Not only is our analysis more general, we also have learned a great deal about the close connection of these processes on finite intervals.

References

LOYNES, R. M. (1962) The stability of a queue with non-independent inter-arrival and service times. *Proc. Camb. Phil. Soc.* **58,** 497–520.

MIYAZAWA, M. (1979) A formal approach to queueing processes in the steady state and their applications. *J. Appl. Prob.* **16,** 332–346.

MIYAZAWA, M. (1983) The derivation of invariance relations in complex queueing systems with stationary inputs. *Adv. Appl. Prob.* **15,** 874–885.

SENGUPTA, B. (1989) An invariance relationship for the $G/G/1$ queue. *Adv. Appl. Prob.* **21,** 956–957.