THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Comparing Batch Delays and Customer Delays

By W. WHITT*

(Manuscript received January 25, 1983)

For a large class of queueing systems in which customers arrive in batches, Halfin (1983) showed that the delay distribution of the last customer in a batch to enter service coincides with the delay distribution of an arbitrary customer when the batch-size distribution is geometric. Halfin's result can be applied to study the performance of complicated communication systems in which messages are divided into packets for transmission. Then packets are customers and the delay of a message is the delay of the last customer in a batch to enter service. If the assumptions are satisfied and if packet delays are easier to analyze, then packet delays can be used to calculate message delays. In this paper, we show that these two delay distributions are stochastically ordered when the batch-size distribution is NBUE or NWUE (new better or worse than used in expectation). The delays of arbitrary customers tend to be less (more) when the batch-size distribution is NBUE (NWUE). In addition to the bounds provided by the stochastic ordering, we also suggest an approximation for the relation between the two expected delays based on known results for the $M^B/G/1$ queue having a batch-Poisson arrival process.

I. INTRODUCTION

Halfin[1] recently showed that for a large class of queueing systems in which customers arrive in batches the delay distribution of the last customer in a batch to enter service equals the delay distribution of an arbitrary customer when the batch size has a geometric distribution. Halfin's result is important when we study the performance of communication systems in which messages are divided into packets for

* Bell Laboratories.

transmission. Then packets are customers and the delay of a message is the delay of the last customer in a batch to enter service. Halfin's result is useful to approximately describe message delays in complicated communication systems, e.g., with contention-resolving schemes such as Carrier Sense Multiple Access (CSMA).[1] In many of these systems message delay is more important but more difficult to analyze than packet delay. Halfin provides conditions under which packet delay results directly yield message delay results.

In this paper we identify conditions on the batch-size distribution under which the delay distribution of the last customer to enter service is stochastically greater than or equal to the delay distribution of an arbitrary customer. We show that it suffices for the batch-size distribution to be NBUE (new better than used in expectation). The ordering is reversed for batch-size distributions that are NWUE (new worse than used in expectation). The batch size B has an NBUE distribution if its mean is greater than all conditional means given tail events, i.e., if $EB \geq E(B - n \mid B \geq n)$ for all n. A distribution is NBUE (NWUE) if it has increasing (decreasing) failure rate. The NBUE and NWUE properties are now quite standard for stochastic comparisons. For further discussion, see Barlow and Proschan,[2] and Whitt.[3]

As Halfin observed, his result about batch arrivals can be viewed as a special case of a discrete analog of Poisson Arrivals See Time Averages (PASTA).[4] In the same way, our stochastic comparison results parallel those for customer-stationary and time-stationary characteristics of queues.[5-7] Random quantities associated with batches (e.g., delays of the last customer in a batch) constitute an embedded sequence in the sequence of random quantities associated with all customers (e.g., the delays of arbitrary customers), just as customer arrival points or departure points constitute embedded sequences in continuous time.

Our approach is similar to that of the East German School (Franken et al.) because we begin in Section II with a stationary version, what we call an "equilibrium batch." In this setting, we easily obtain a representation of the expected average associated with an arbitrary customer that makes the desired comparisons and Halfin's result transparent. The representation involves the stationary-excess distribution of the batch-size distribution. (From the theory of stationary point processes, as contained in Franken et al., this can also be viewed as a consequence of the Palm theory.)

The relation between what an arbitrary customer sees and what the last customer in a batch sees can be explained in part as follows. One customer in each batch is the last customer in the batch, but there are more customers in big batches than in small batches. Hence, the distribution of the batch size containing a last customer is the ordinary

batch-size distribution, but the distribution of the position in a batch of an arbitrary customer is the batch-size stationary-excess distribution (see Section III.2 of Cohen,[8] Burke,[9] and Section 5.10 of Cooper[10]). Hence, the relation between the two kinds of delays reduces to the relation between the batch-size distribution and its associated stationary-excess distribution.

These stochastic comparison results are only qualitative. With the methods we use, we are unable to obtain related quantitative results. Moreover, the difference between the expected delays obviously will depend on the context, whereas the qualitative results here generally hold true. Following Halfin's example, we make very few assumptions for our stochastic comparisons. However, in Section III we examine quantitative results by considering the special case of the $M^B/G/1$ queueing model, which has a batch-Poisson arrival process. We get some idea about what to expect in more complicated situations by examining existing quantitative results for this special case. We use these results to obtain an approximate quantitative relationship between expected batch delays and expected customer delays as a function of the first two moments of the batch-size distribution.

In Section IV we briefly show how our equilibrium batch can be viewed as a time-average limit. Throughout we talk about batch arrivals to a queue, but as in Wolff[4] the results apply more generally. The model need not be a queue, and for a queueing model the process need not be an arrival process; for an arrival process the customers in a batch need not arrive together. The batch is simply a label assigned to random variables, as we explain in Section II.

II. THE EQUILIBRIUM BATCH

We consider a batch in equilibrium, defined in terms of a positive-integer-valued random variable B and a sequence of random variables $\{X_k, k \geq 1\}$, all defined on a common probability space. The variable B represents the size of the batch and the variable X_j is a random quantity associated with the jth customer in the batch. For example, X_j might be a function of the delay of the jth customer to enter service. We are only interested in X_j for $j \leq B$. In fact, we can consider X_j defined only on the subset $\{B \geq j\}$. Hence, we can speak of the basic data as the random vector $(B, X_1, \cdots, X_B)$.

The variables $X_1, X_2, \cdots$ are typically dependent. Our *basic assumption* is that the events $\{B = j\}$ for $j \geq k$ are independent of the random vector $(X_1, \cdots, X_k)$ for $k \geq 1$. This corresponds to the Lack of Anticipation Assumption (LAA) in Wolff.[4]

Let the probability mass function (pmf) of B be defined as

$$p_n = P(B = n), \qquad n = 1, 2, \cdots, \tag{1}$$

and let p_n^* be the *stationary-excess* pmf associated with p_n, defined by

$$p_n^* = \bar{p}_n \Big/ \sum_{k=1}^{\infty} \bar{p}_k, \qquad n \geq 1, \tag{2}$$

where

$$\bar{p}_n = \sum_{k=n}^{\infty} p_k, \qquad n \geq 1, \tag{3}$$

and

$$\sum_{k=1}^{\infty} \bar{p}_k = \sum_{k=1}^{\infty} k p_k = EB < \infty. \tag{4}$$

We are interested in the relationship between the random quantities associated with an *arbitrary* customer and the *last* customer in a batch so we can compare the expected values in these two cases. We interpret the expected value associated with an arbitrary customer as the expected value for all customers in the batch divided by the expected number of customers in a batch. (A way to justify this interpretation is described in Section IV). Because of our basic assumptions, these quantities are

$$A = \frac{\sum_{n=1}^{\infty} p_n (EX_1 + \cdots + EX_n)}{\sum_{n=1}^{\infty} n p_n} \tag{5}$$

and

$$L = EX_B = \sum_{n=1}^{\infty} p_n EX_n. \tag{6}$$

The desired relationships between A and L are derived from the following alternate representation for A.

Theorem 1: $A = \sum_{n=1}^{\infty} p_n^* EX_n$.

Proof: Change the order of summation in (5) and apply (2) to obtain

$$A = \sum_{n=1}^{\infty} EX_n \left(\frac{\sum_{k=n}^{\infty} p_k}{\sum_{k=1}^{\infty} k p_k} \right) = \sum_{n=1}^{\infty} EX_n \left(\frac{\bar{p}_n}{\sum_{k=1}^{\infty} \bar{p}_k} \right) = \sum_{n=1}^{\infty} EX_n p_n^*.$$

Corollary 1: (*Halfin*[1]) *$A = L$ for all $\{EX_n\}$ if and only if $p_n = p_n^*$ for all n or, equivalently, if p_n is geometric, i.e.,*

$$p_n = (1 - p)p^{n-1}, \qquad n = 1, 2, \cdots \tag{7}$$

for some p, $0 \leq p \leq 1$.

Proof: Sufficiency is immediate. It is well known that p_n is geometric if and only if $p_n = p_n^*$ for all n (e.g., see Corollary 3.3 of Whitt[3]). Necessity is almost as easy: Choose different sequences $\{EX_n\}$, e.g., $EX_n = 1$ if $n = k$, and 0 otherwise.

Remark: Theorem 1 and its corollaries apply immediately to distributions as well as means. For example, our original sequence X_n can be replaced by $f(X_n)$ where $f(x) = I_{(-\infty,x)}$, so that A is the expected proportion of customers for which $X_n \le x$ and $L = P(X_B \le x)$.

We now establish inequalities between A and L as a function of the shape of the batch-size pmf p_n. These follow immediately from known stochastic-order relations between p_n and p_n^*.

For two pmf's p_n^1 and p_n^2 on the positive integers, we define stochastic-order relations $p_n^1 \le_{st} p_n^2$ ($p_n^1 \le_{ic} p_n^2$) to hold if

$$\sum_{k=1}^{\infty} f(k)p_k^1 \le \sum_{k=1}^{\infty} f(k)p_k^2 \tag{8}$$

for all nondecreasing (nondecreasing and convex) real-valued functions f on the positive integers for which the sums converge.

A pmf p_n is NBUE (new better than used in expectation) if

$$EB = \sum_{k=1}^{\infty} \bar{p}_k \ge \sum_{k=n}^{\infty} \bar{p}_k/\bar{p}_n = E(B - n \mid B \ge n), \qquad n \ge 1, \tag{9}$$

and NWUE with the inequality in (9) reversed.

From Theorem 1, we obtain:

Corollary 2: $A \le L$ for all nondecreasing sequences $\{EX_n\}$ if and only if $p_n^* \le_{st} p_n$ or, equivalently, if p_n is NBUE.

Proof: From (6), (8), and Theorem 1, $A \le L$ for all nondecreasing $\{EX_n\}$ if and only if $p_n^* \le_{st} p_n$. It is well known that $p_n^* \le_{st} p_n$ if and only if p_n is NBUE [e.g., see Theorem 3.2 (iii) of Whitt[3]]. For necessity, choose $EX_n = 1$ for $n \ge k$, and 0 otherwise.

Remarks: (1) Corollary 2 remains valid with $p_n^* \ge_{st} p_n$ instead of $\le_{st}$, which is equivalent to p being NWUE, if either $A \ge L$ or $\{EX_n\}$ is nonincreasing (but not both). (2) For Halfin's problem involving delays, note that the delays of the successive customers in any batch to begin service are nondecreasing with probability one.

We obtain another corollary using the stochastic-order relation $\le_{ic}$ defined in (8).

Corollary 3: $A \le L$ for all nondecreasing convex sequences $\{EX_n\}$ if and only if $p_n^* \le_{ic} p_n$ or, equivalently, if p_n new is better than p_n^* used in expectation, i.e., if

$$\sum_{k=1}^{\infty} \bar{p}_k \ge \sum_{k=n}^{\infty} \bar{p}_k^*/\bar{p}_n^*, \qquad n \ge 1.$$

Proof: Follow the proof of Corollary 2 using the convexity to treat $\leq_{ic}$. The characterization of $p_n^* \leq_{ic} p_n$ follows easily from the fact that $p_n^1 \leq_{ic} p_n^2$ if and only if $\sum_{k=n}^{\infty} \bar{p}_k^1 \leq \sum_{k=n}^{\infty} \bar{p}_k^2$; see Definition 3.1(iv) and Theorem 3.2(iv) of Whitt.[3]

III. The $M^B/G/1$ QUEUE

To illustrate the qualitative results and obtain some related quantitative results, we now consider the special case of the $M^B/G/1$ queue having a batch-Poisson arrival process (see Section 5.10 of Cooper[10]). Let W_F, W_A, and W_L be the equilibrium delay (waiting time before beginning service) of the first customer in a batch, an arbitrary customer, and the last customer in a batch. Let τ and σ^2 be the mean and variance of the service time; let m and $\hat{\sigma}^2$ be the mean and variance of the batch size B; and let B^* have the batch-size stationary-excess distribution p_n^* in (2), which has mean

$$EB^* = \sum_{n=1}^{\infty} np_n^* = (\hat{\sigma}^2 + m^2 + m)/2m. \tag{10}$$

Let λ be the rate of the Poisson process and, for stability, assume that $\lambda m\tau < 1$.

From Cooper, we obtain

$$E(W_F) = \frac{\lambda m^2\tau^2}{2(1 - \lambda m\tau)}\left(1 + \frac{m\sigma^2 + \tau^2\hat{\sigma}^2}{m^2\tau^2}\right). \tag{11}$$

From Theorem 1 and (10), we obtain

$$\begin{aligned} E(W_A) &= E(W_F) + \sum_{n=1}^{\infty} p_n^*(n-1)\tau \\ &= E(W_F) + [E(B^*) - 1]\tau \\ &= E(W_F) + \frac{(m-1)\tau}{2} + \frac{\hat{\sigma}^2\tau}{2m}. \end{aligned} \tag{12}$$

From (6), we obtain

$$\begin{aligned} E(W_L) &= E(W_F) + (EB - 1)\tau \\ &= E(W_F) + (m-1)\tau. \end{aligned} \tag{13}$$

Let c_B^2 be the squared coefficient of variation of the batch-size B, i.e., the variance of B divided by the square of the mean. We can rewrite (12) as

$$E(W_A) = E(W_F) + [m(c_B^2 + 1) - 1]\tau/2, \tag{14}$$

so that $EW_A \leq EW_L$ if and only if

$$mc_B^2 + 1 \leq m, \tag{15}$$

which is consistent with Corollaries 1 through 3 in Section II because $m = 1/p$ and $c_B^2 = 1 - p$ for the geometric distribution in (7).

In applying Theorem 1, we should observe that the lack of anticipation (LAA) assumption holds for the $M^B/G/1$ queue. Moreover, the Poisson property is only used to get (11); (12) and (13) remain valid provided that the LAA assumption holds. For example, the LAA assumption holds if the intervals between batch arrivals are a stationary sequence independent of the successive batch sizes.

Formulas (11) through (14) suggest an approximation for more general systems:

$$\frac{E(W_A) - E(W_F)}{E(W_L) - E(W_F)} = \frac{m(c_B^2 + 1) - 1}{2(m - 1)}. \tag{16}$$

IV. LIMITING AVERAGES

Instead of using the framework defined in Section II, we could also begin with a sequence of random vectors $\{(B_k, X_{k1}, \cdots, X_{kB_k}), k \geq 1\}$, where B_k represents the size of the kth batch, indexed in order of arrival, and X_{kj} represents the random quantity of interest (delay, etc.) associated with the jth customer to enter service in the kth batch. Let B_k take values on the positive integers and let X_{kj} be nonnegative for each k and j. Let Y_k be the random quantity associated with the kth customer indexed, first according to the batch and then according to the order of entering service, defined by

$$Y_k = X_{jm}, \qquad B_1 + \cdots + B_{j-1} + m = k \leq B_1 + \cdots + B_j$$

for $k \geq 1$.

The limiting average value of X_{kj} over *all* customers is naturally defined by

$$\bar{A} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} Y_k}{n}. \tag{17}$$

We shall work with the related quantity $\hat{A}$ defined by

$$\hat{A} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \sum_{j=1}^{B_k} X_{kj}}{\sum_{k=1}^{n} B_k}. \tag{18}$$

In most situations, the limits $\bar{A}$ and A exist and are equal.

We are interested in the relation between A and the limiting average value of X_{kj} over the *last* customer in each batch, defined by

$$\hat{L} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} X_{kB_k}}{n}. \tag{19}$$

We assume that there exists a random vector $(B, X_1, \cdots, X_B)$ such that

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} B_k}{n} = EB < \infty \tag{20}$$

and

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \sum_{j=1}^{B_k} X_{kj}}{n} = E \sum_{j=1}^{B} X_j < \infty, \tag{21}$$

so that

$$\hat{A} = \frac{E \sum_{j=1}^{B} X_j}{EB}. \tag{22}$$

We also assume that

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} X_{kj} I_{\{B_k = j\}}}{n} = P(B = j)EX_j. \tag{23}$$

To obtain (23) it is natural to assume that the basic independence assumption holds for each k, i.e., $\{B_k = j\}$ is independent of $(X_{k1}, \cdots, X_{kn})$ for all $j, j \geq n$. From (23), we have

$$\hat{L} = EX_B. \tag{24}$$

For eqs. (20) through (24) to be valid, in addition to the basic independence assumption, it suffices for the basic sequence $\{(B_k, X_{k1}, \cdots, X_{kB_k}), k \geq 1\}$ to be stationary and ergodic.

With (22) and (24) we obtain the framework defined in Section II.

V. ACKNOWLEDGMENTS

I am grateful to Shlomo Halfin for introducing me to this problem and to Moshe Segal for making helpful comments.

REFERENCES

1. S. Halfin, "Batch Delays Versus Customer Delays," B.S.T.J., this issue.
2. R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, New York: Holt, Rinehart and Winston, 1975.

3. W. Whitt, unpublished work.
4. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–31.
5. D. König and V. Schmidt, "Stochastic Inequalities Between Customer-Stationary and Time-Stationary Characteristics of Queueing Systems With Point Processes," J. Appl. Prob., *17*, No. 3 (September 1980), pp. 768–77.
6. P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*, Berlin: Akademie-Verlag, 1981.
7. S. Niu, unpublished work.
8. J. W. Cohen, *The Single Server Queue*, Amsterdam: North Holland, 1969.
9. P. J. Burke, "Delays in Single-Server Queues With Batch Input," Oper. Res., *23*, No. 4 (July–August 1975), pp. 830–3.
10. R. B. Cooper, *Introduction to Queueing Theory*, Second Edition, New York: North Holland, 1981.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; Bell Laboratories, 1977—. At Yale University, from 1973 through 1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At Bell Laboratories he is in the Operations Research Department. His work focuses on stochastic processes and stochastic models in operations research.