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

Batch Delays Versus Customer Delays

By S. HALFIN*

(Manuscript received January 6, 1982)

For a large class of contention schemes with messages transmitted by subdividing them into packets, we show that the delay distribution of a message is the same as that for an individual packet. We conclude this from analyzing queueing systems with batch arrivals, where batch sizes have a geometric distribution and the queue discipline is indifferent to batch sizes and service times. There we prove that the customer (packet) delay distribution is the same as the batch (message) delay distribution, where delay is defined to be the delay of the *last served* customer in the batch. The proof is based on the discrete-time analog of the Poisson Arrivals See Time Averages (PASTA) theorem. We conclude that, in many cases, we can obtain message delays by calculating or measuring the packet delays, which is usually an easier task.

I. INTRODUCTION

Consider a queueing system with batch arrivals. We define the delay of batch to be the maximal delay of the customers in the batch. For example, consider the case where data messages arrive at a node of a network and await transmission. Each message is subdivided into packets that may be transmitted individually. Here the customers correspond to the packets and the batch to a message, and it is natural to say that the message is delayed as long as at least one of its packets is delayed.

Next, assume that the number of customers in a batch has a geometric distribution:

$$\Pr(\text{batch size} = n) = p(1-p)^{n-1}, \qquad n = 1, 2, \cdots.$$

* Bell Laboratories.

In our example, if the message lengths (in bits) are exponentially distributed, and each message m is "chopped" into $i(m)$ packets whose lengths (in bits) are independent and identically distributed and independent of the message length, then $i(m)$ has a geometric distribution. [Note that the $i(m)$th packet usually will not be full.]

We prove that for a wide class of such systems the batch delay in equilibrium has the same distribution as the individual customer delay. Here delay is the time from arrival to start of service. In Section II we discuss the various queue disciplines for which this result holds. A discete-time analog of the Poisson Arrivals See Time Averages (PASTA) theorem is introduced in Section III. The main result is stated and proved in Section IV. Section V contains some additional comments, and conclusions are stated in Section VI.

II. QUEUE DISCIPLINES

The stated result does not hold for all queue disciplines. Consider the case where customers are selected for service independent of their service times. Then it is well known that the expected delay of a customer is independent of the queue discipline.[1] On the other hand, the expected batch delay may depend on the discipline. For instance, if one always chooses the next customer from a batch with the smallest number of remaining customers, the expected batch delay will be smaller than if the next batch is chosen randomly, or in a preassigned order. This can be proved by arguments similar to those showing that giving preferential treatment to customers with small service time reduces the expected waiting time.[2] Thus, to prove equality of the delay distributions, such disciplines must be excluded.

Next, we characterize disciplines for which the result holds:

Definition: A queueing discipline will be called *impartial* if it selects customers independently of their services times, and selects batches independently of their sizes.

The following are examples of impartial disciplines:

- First in, first out (FIFO) for batches and for customers within batches.
- Last in, first out (LIFO) for batches and for customers within batches.
- Random choice of a batch, and then a random choice of a customer from that batch (but not a random choice among all waiting customers because this will favor large batches).
- Random choice of a batch, and then serving all the customers of that batch in FIFO, LIFO, or random order.
- Contention schemes: each batch chooses a candidate customer for next service (independent of its service time) and the contention

between the candidates is resolved independent of the sizes of their respective parent batches.

The last family includes disciplines that are applicable to the case of transmitting messages by subdividing them into packets. Here the candidate packet for transmission in each message is the first packet not yet transmitted. Examples of relevant contention-resolving methods are (1) the round robin (token) scheme, and (2) Carrier Sense Multiple Access (CSMA) scheme.[3-5]

III. DISCRETE-TIME ANALOG OF PASTA

The following result will be needed later. Let X_n, $n = 1, 2, \cdots$ be a sequence of random variables, and let B be a set in the value space of X. Let U_n be the indicator function of the event $X_n \in B$, for all n. Let A_n, $n = 1, 2, \cdots$ be a sequence of i.i.d. Bernoulli random variables defined on the same probability space as the X_n's.

$$\text{Let} \quad V_n = n^{-1} \sum_{i=1}^{n} U_i, \quad Y_n = \sum_{i=1}^{n} U_i A_i, \quad \text{and} \quad Z_n = Y_n \left(\sum_{i=1}^{n} A_n \right)^{-1}.$$

Then the following hold:

Theorem 1: If for every n the set of random variables $X_1, X_2, \cdots, X_n$ is independent of the set $A_n, A_{n+1}, \cdots$ then: $V_n \to V$ w.p. 1 if and only if $Z_n \to V$ w.p. 1, as $n \to \infty$.

Remark: This theorem is a discrete-time analog of the PASTA theorem,[6] which is stated in terms of a continuous time stochastic process $X(t)$, and a Poisson process $A(t)$. The assumption in the theorem is called by Wolff 'Lack of Anticipation Assumption' (LAA). The name PASTA comes from the special case where $A(t)$ is an arrival process to a queue, and $X(t)$ is the number of customers in the system at time t. In that case the theorem states that the long-term proportion of time for which $X(t) = m$ (arbitrary nonnegative integer) is equal to the long-term proportion of arriving customers that find the system in state m.

Proof: Wolff's Lemma 1, Lemma 2, and Theorem 1[6] can be easily stated and verified in the discrete-time case, and thus the conclusion holds.

IV. THE RESULT AND ITS PROOF

Theorem 2: Given a queueing system with batch arrivals where:

1. *The batch sizes are i.i.d. with a geometric distribution, and independent of the arrival times process*
2. *The service times are independent of the arrival times and batch sizes*

3. *The queue discipline is impartial, then the customer delay distribution is equal to the batch delay distribution.*

Remark: We interpret the conclusion of the theorem in a time average sense. Thus, for any $x \geq 0$ we compare the proportions of customers and batches that are delayed x time units or less. We prove that if one of these proportions has a long-term limit, so does the other, and they are equal. In the common case, when the queueing system is ergodic, the conclusion can be also interpreted in terms of delay distributions of individual customers and batches.

Proof: Let $X_1, X_2, \cdots$ be the sequence of delays of customers arranged in the order in which they go into service. Let $x \geq 0$ be fixed, and let $B = [0, x]$. Let $A_n = 1$, if the nth customer in the above order is the last to go to service in its batch, and $A_n = 0$ otherwise. Next, observe that the LAA assumption holds for this setup because for any n, $X_1, \cdots, X_n$ are determined by the arrival process, service times of the first $n - 1$ served customers, and queue discipline, all of which are independent of the batch size of the nth customer. (Note that if the discipline is not blind to service time, the above service times could depend on the batch sizes. Also note that the delays do depend in general on the *number* of present batches, and thus A_n and X_{n+1} would typically be dependent. For instance, if $A_n = 1$, the probability that no customer remains waiting, implying $X_{n+1} = 0$, increases.) The proof is now completed by applying Theorem 1.

V. FURTHER COMMENTS

The assumptions of Theorem 2 are quite weak. The theorem is true for a queue with many servers, even with different service rates, as long as the assignment of customers to servers is again independent of the present batch sizes.

We do not require that a service begins immediately when a server becomes free, even if customers are waiting. There may be a "dead period", as is the case in some contention schemes. However, the length of this dead period has to be independent of the present batch sizes.

If we want to derive a similar theorem for *time in system* (rather than delay), then it will not hold in general in the multiserver case. For example, it will not be true for an infinite number of servers, and nonconstant service time. (Ward Whitt[7] provided this example.)

Although the last customer to enter service in a given batch has the same distribution of time in system as a typical customer, its time in system may be different from the batch's time in system. This is so because, in the multiserver case, when that customer completes service, other customers of its batch may still be in service. If we apply

Theorem 2 to the setup where we order the customers by the time they leave the system and let the X_n's be the corresponding times in system, the application will fail because the LAA assumption will no longer be valid.

VI. CONCLUSION

We have shown that for many queueing systems with batch arrivals having geometrically distributed sizes, the individual customers experience the same delay distribution as the batches themselves. This seems to affect the case of transmitting messages by packets, since in many cases it is possible to obtain packet delays, while from a performance point of view one is more interested in the message delays.

VII. ACKNOWLEDGMENT

I would like to thank Ward Whitt for his useful comments and suggestions.

REFERENCES

1. L. Kleinrock, *Queueing Systems*, Vol. 2, New York: Wiley, 1976, p. 113.
2. R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*, New York: Addison-Wesley, 1967.
3. L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part I—Carrier Sense Multiple Access Modes and Their Throughput—Delay Characteristics," IEEE Trans. Commun., *23* (1975), pp. 1400–16.
4. F. A. Tobagi and V. B. Hunt, "Performance Analysis of Carrier Sense Multiple Access with Collision Detection," Proc. LACN Symp., May 1979, pp. 217–44.
5. D. P. Heyman, "An Analysis of the Carrier Sense Multiple Access Protocol," B.S.T.J., *61*, No. 8 (October 1982), pp. 2023–51.
6. R. W. Wolff, "Poisson Arrivals See Time Averages," Op. Res., *30*, No. 2 (1982), pp. 223–31.
7. W. Whitt, private communication.

AUTHOR

Shlomo Halfin, M.Sc., 1958, and Ph.D., 1962 (Mathematics), The Hebrew University of Jerusalem; Bell Laboratories, 1968—. Mr. Halfin's areas of interest are stochastic processes and mathematical programming. He currently works on applications of these methods to communication networks. Member, SIAM, ORSA.