Analysis of a Finite Storage System with Batch Input Arising out of Message Packetization

DAVID R. MANFIELD, MEMBER, IEEE, AND P. TRAN-GIA, MEMBER, IEEE

Abstract—**A Markovian queueing model with finite waiting space is developed for the communications controller which buffers the flow of packets from a host computer to its associated packet switch. The segmentation of messages into packets by the host is modeled by a batch input to the communications controller. The probabilities of state are determined by numerical recursion and subsequently used in expressions developed for the blocking probabilities and waiting-time distribution as a function of two proposed batch acceptance strategies. Representative numerical results as would be useful in the dimensioning and performance analysis of the communications controller are presented. A more general non-Markovian model is investigated by means of simulation, showing that the initial Markovian model is very accurate in determining the system performance.**

I. INTRODUCTION

INPUT traffic to the nodes of a packet-switching network is directed from associated host computers located externally to the network. The high-speed host receives messages from a large number of terminals and remote concentrators (multiplexers), and these messages are subjected to a packetizing function in the host before being sent to the associated packet switch for transmission to the destination through the actual packet-switching network. The packetizing function leads to the situation where batches of packets are presented simultaneously to the communications controller (Fig. 1) for sending to the packet switch. Moreover, the nature of the operation of the system is such that the arrival of messages at the host constitutes a random process and so the packet stream from the host is buffered by the communications controller. In physical realizations this buffer is finite and, due to the random process of arrivals, it will experience conditions of overflow. It is of fundamental interest to analyze this buffer both for dimensioning purposes and for determining its performance under various traffic conditions. With this motivation, the communications controller providing the interface from the host to the packet-switching network is modeled as a finite queue subjected to a batch input process.

Some related problems have been considered before in the literature, but set in different contexts. In [1]-[3], a transmission buffer for a statistical multiplexer is modeled as a batch arrival queue with constant, synchronous output. Here the input batches are numbers of characters in arriving messages. Although the intent of this work is to examine finite buffer lengths, the cases considered are all those when the probability of blocking is so small that the buffer is essentially infinite. This allows the use of an approximate method to find the (very small) buffer overflow probability. The state probabilities are found using a numerical technique involving fast Fourier transforms. A number of works have included batch arrivals in the context of the modeling of common-control switching systems [4]-[7], but in each case, the buffer sizes are assumed to be infinite for the sake of analysis. The motivation for batch inputs in these systems is generally the accumulation of customers between clocked arrival times. In [15], the input buffer to a packet switch is analyzed for batch arrivals arising from the packetizing of incoming messages, but here, again, the treatment is for infinite buffer size.

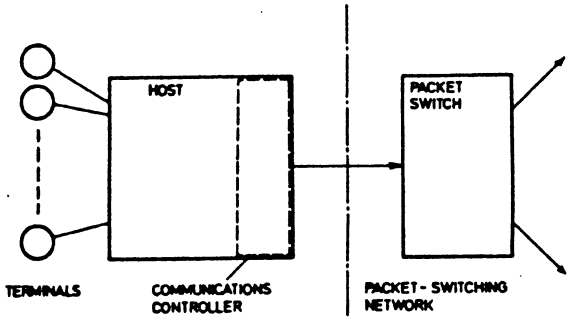


Fig. 1. Host-packet switch interface.

A number of theoretical studies for infinite, batch arrival queues have been performed [8]-[12] with results of varying degrees of complexity. In [13], [14] are very mathematical treatments of finite waiting systems with batch arrivals, but it is not easily seen how the results may be used to practical advantage. Because of the mathematical complexity of finite, batch arrival queues, the work presented here begins with the least complicated model to illustrate the ideas, namely, a finite Markovian queue with batch arrivals. The rationalization of this model is given in the next section, with a brief discussion of the analytical difficulties. In Sections III and IV are given, respectively, the development of the buffer overflow probabilities and waiting-time distribution for packets, and the numerical results for some representative examples. Finally, in Section V, the relation of these results to those obtained from more general models by simulation is discussed.

II. MODELING

We proceed in a straightforward way to the modeling of the communications controller which operates at the interface

Paper approved by the Editor for Computer Communication of the IEEE Communications Society for publication without oral presentation. Manuscript received December 8, 1980. This work was supported by the Alexander von Humboldt Foundation.

D. R. Manfield was with the Department of Communications, University of Siegen, 5900 Siegen-1, West Germany. He is now with Bell Northern Research, Ottawa, Ont., Canada K1Y 4H7.

P. Tran-Gia is with the Department of Communications, University of Siegen, 5900 Siegen-1, West Germany.

0090-6778/82/0300-0456$00.75 © 1982 IEEE

between the host and the packet switch. The messages arriving at the host from the terminals are referred to as the arriving *batches*, and the packets subsequently produced by the host's packetizing function are referred to as the *customers* which constitute a batch. Each customer occupies one waiting place in the buffer of the communications controller. The following two initial assumptions are made.

Assumption 1) The instants of batch arrivals constitute a stationary Poisson process with rate λ.

Assumption 2) The service times of individual customers are independent, identically distributed random variables with negative exponential distribution, mean $1/\mu$.

The first assumption is based on the observation that the incoming message stream is the superposition of offered traffics from a large number of different terminals and users, and it has a strong basis in traffic theory. The second assumption is not so easily justified, and is strongly related to how the server of the communications controller is visualized. The service time is the time required to transmit a packet from the host to the packet switch, and since the packets will generally be of fixed length, there is an initial argument for fixed (deterministic) service times as in [1], [11], [15]. However, it is noted that the transmission of the packets to the packet switch is under the control of a host-packet switch protocol [16] and the transmission time can be affected by a number of factors other than the packet length, such as the need for retransmissions arising from errors, and the temporary suspension of packet acceptance at the packet switch as would result from a full receive buffer in the packet switch. In this way, it is seen that the service time for packets in the communications controller is in fact an "effective" service time for the transmission of packets to the packet switch, and provides the basis for the assumption of service times as random variables with exponential distribution. The service time distribution will be discussed further in Section V. It should be noted that the following analysis will apply to a general batch-size distribution, although for the numerical results in Section IV, particular forms of the batch-size distribution (e.g., geometric, Poisson) will be used.

Thus, it is seen that the model for the communications controller is a queueing system of the type $M^{[x]}/M/1-s$ where s denotes the number of waiting places. This is the simplest type of finite, batch queue, but although it is Markovian, it does not seem to be possible to obtain closed form expressions for the state probabilities, or even the generating function of these probabilities. Furthermore, the finite waiting room excludes the possibility of using the so-called "super-customer" approach [9] for dealing with the batch arrival process. In the next section, the state probabilities are found by a numerical recursion method, and form the basis for the subsequent calculations of blocking probabilities and waiting-time distribution.

The modeling is not yet complete, since it is necessary to also consider the operation of the system when an arriving batch is larger in size than the number of unoccupied waiting places. To account for different blocking modes, the following two *batch acceptance strategies* are defined.

Strategy 1) An arriving batch larger in size than the number of available free queue positions is totally rejected.

Strategy 2) An arriving batch larger in size than the number of available free waiting places fills the free positions and the remaining customers of the batch are lost.

These strategies will be referred to, respectively, as the *whole batch* and *part batch* acceptance strategies. Finally, the following is assumed.

Assumption 3) Blocked customers depart immediately and no longer affect the system.

In the context of the communications controller model, this is clearly only a starting assumption, since blocked packets cannot be expected to simply go away. In real systems, it is most likely that there would be some form of large secondary storage medium, where packets blocked from the "send" buffer of the communications controller would be kept for later transmission. It is felt that for the analysis of such systems with secondary backup storage the finite model treated here is an essential first step. Also, under certain network load conditions, e.g., temporary overload, the size of the send buffer of the communications controller may be made artificially small to prevent too many new packets entering the packet-switching network. In this case, it is necessary to know the (large) buffer overflow probability in order to determine how many packets are being routed to the backup storage.

III. ANALYSIS

A. State Probabilities

Since the queueing system for the model is Markovian, the state probabilities are described by a set of Chapman-Kolmogorov difference equations in the steady state, using standard techniques [9]. However, since the arrivals occur in batches, the state probabilities are not confined to simple one-state transitions. The first requirement is to solve for the state probabilities, which we do for the $M^{[x]}/M/m-s$ system since it involves no extra difficulty. Let

$$N = m + s \qquad \text{(system size)}$$

$$P_n = \Pr\{\text{arriving batch encounters } n \text{ in system}\}$$

$$g_i = \Pr\{\text{arbitrary batch is of size } i;\ i \geq 1\}.$$

For the purposes of illustration, we consider three forms for the distribution of the $\{g_i\}$, namely

$$g_i = \begin{cases} \theta^{i-1}(1-\theta), & i \geq 1 \qquad \text{(geometric)} \\ \dfrac{\theta^{i-1}}{(i-1)!}e^{-\theta}, & i \geq 1 \qquad \text{(shifted Poisson)} \\ \dfrac{1}{N_{\max}}, & 1 \leq i \leq N_{\max} \quad \text{(uniform).} \end{cases}$$

By inspection, the following set of Chapman-Kolmogorov equations may be seen to hold.

Strategy 1)

$$0 = -\lambda \sum_{i=1}^{N} g_i P_0 + \mu P_1$$

$$0 = -\left(\lambda \sum_{i=1}^{N-k} g_i + k\mu\right) P_k + (k+1)\mu P_{k+1} + \lambda \sum_{i=0}^{k-1} g_{k-i} P_i, \qquad (0 < k < m)$$

$$0 = -\left(\lambda \sum_{i=1}^{N-k} g_i + m\mu\right) P_k + m\mu P_{k+1} + \lambda \sum_{i=0}^{k-1} g_{k-i} P_i, \qquad (m \leq k < N)$$

$$0 = -\mu P_N + \lambda \sum_{i=0}^{N-1} P_i g_{N-i}. \qquad (1)$$

Strategy 2)

$$0 = -\lambda P_0 + \mu P_1$$

$$0 = -(\lambda + k\mu) P_k + (k+1)\mu P_{k+1} + \lambda \sum_{i=0}^{k-1} g_{k-i} P_i, \qquad (0 < k < m)$$

$$0 = -(\lambda + m\mu) P_k + m\mu P_{k+1} + \lambda \sum_{i=0}^{k-1} g_{k-i} P_i, \qquad (m \leq k < N)$$

$$0 = -m\mu P_N + \lambda \sum_{i=0}^{N-1} P_i \sum_{j=N-i}^{\infty} g_j. \qquad (2)$$

From these equations, it may be seen that for either strategy, the state transition matrices are triangular, and hence the state probabilities $\{P_n\}$ can be found by the numerical procedure of direct forward recursion by assuming a value for P_0 and at the end utilizing the normalizing relation $\Sigma P_n = 1$. These state probabilities form the basic requirement for the calculations of blocking probabilities and waiting-time distributions which now follow.

B. Blocking Probabilities

It is necessary to distinguish between the blocking probability for an arbitrary batch and the blocking probability of an arbitrary customer, since in general they are not the same. It is clear from the nature of the system that any state can be a blocking state. The batch blocking probability under Strategy 2, the part batch acceptance mechanism, is not uniquely defined and, for our purposes, we define it to be the probability that a batch is either partly or wholly rejected. The batch blocking probability is found by considering an arbitrary "test" batch and conditioning on the system state found by this batch. The resultant expression is the same for each acceptance strategy and it is

$$B_{\text{batch}} = \sum_{k=0}^{N} P_k \sum_{i=N-k+1}^{\infty} g_i. \qquad (3)$$

Given the batch finds k in the system, the batch will be blocked if it is bigger than $N - k$ in size. Although (3) is independent of acceptance strategy, the blocking probability will nonetheless depend on the strategy by virtue of the differing sets of state probabilities.

The blocking probability for individual customers is not quite so easily obtained. On consideration of an arbitrary "test" customer, it is known from [17] that the batch containing this test customer is of size i given by

$$\Pr\{\text{arbitrary customer arrives in batch of size } i\} = ig_i/\text{MBS} \qquad (4)$$

where MBS is the mean batch size. That is, the test customer is more likely to arrive in a longer batch than a shorter one. Conditioning on the system state found by the batch containing the test customer, and noting that the test customer occupies any position in its batch with equal probability, we obtain the following.

Strategy 1)

$$B_{\text{customer}} = \frac{1}{\text{MBS}} \sum_{k=0}^{N} P_k \sum_{i=N-k+1}^{\infty} ig_i \qquad (5)$$

Strategy 2)

$$B_{\text{customer}} = \frac{1}{\text{MBS}} \sum_{k=0}^{N} P_k \sum_{i=N-k+1}^{\infty} (i - N + k) g_i. \qquad (6)$$

Equation (5) for Strategy 1 follows from the fact that when the batch containing the test customer is larger than the available waiting space, all the customers of this batch are lost. Under Strategy 2, when the batch is too large for the available space, the test customer occupies a position in the rejected portion of the batch with probability $(i - N + k)/i$, given the batch found state k, and hence, (6) follows. Equations (5) and (6) have an alternative derivation in terms of the ratio of number of lost customers to number of arriving customers over a long time interval. In this case, the numerators of (5) and (6) are thought of as expected lengths of blocked batches (or part batches), and the denominators as the expected lengths of all arriving batches.

C. Waiting Time Distribution

To find the waiting time distribution function, some further results of [17] are drawn upon, and again the idea of an arbitrary test customer is used. The queue discipline is assumed to be first-come, first-served. The test customer can suffer one of three fates. It can be blocked on arrival, it can find a free server and, hence, go directly into service, or it can join the waiting line. In the first two cases, the test customer is assigned a zero waiting time. The waiting time is composed of two separate components; first, the waiting time suffered by the first-served customer of a batch, that is, the "residual" waiting time caused by customers already in the system, and second, the waiting time suffered by the test customer caused by its position within its own batch. Both of these

components will be a convolution of a number of exponential distributions, and the overall waiting-time distribution is a convolution of the two components of waiting. The conditional waiting distribution for customers which actually have to wait is then obtained by a process of normalization.

The residual waiting time component depends on the number of customers in the system and whether all the servers are busy. If one or more servers are free then at least part of the batch goes directly into service and the residual waiting time is zero. If no server is free, then the residual waiting time is the time for all the already waiting customers to be served. For the second component of waiting caused by the position of the test customer within its batch, the test customer in the jth position (all positions equally probable) means it has to wait for the $j-1$ customers ahead of it to be served. Using all these facts, and recalling (4), the following equations are derived.

Define

$$\bar{W}(t) = \Pr\{\text{test customer waits time } \tau > t \mid \text{customer is accepted}\}.$$

Strategy 1)

$$\bar{W}(t) = \left[\sum_{k=0}^{m-1} P_k \sum_{i=m-k+1}^{N-k} g_i \sum_{j=0}^{i-m+k-1} \sum_{l=0}^{j} \frac{(m\mu t)^l}{l!} e^{-m\mu t} + \sum_{k=m}^{N-1} P_k \sum_{i=1}^{N-k} g_i \sum_{j=0}^{i-1} \sum_{l=0}^{j+k-m} \frac{(m\mu t)^l}{l!} e^{-m\mu t}\right] \cdot \left[\sum_{k=0}^{N-1} P_k \sum_{i=1}^{N-k} i g_i\right]^{-1}. \quad (7)$$

Strategy 2)

$$\bar{W}(t) = \left[\sum_{k=0}^{m-1} P_k \left[\sum_{i=m-k+1}^{N-k} g_i \sum_{j=0}^{i-m+k-1} \sum_{l=0}^{j} \frac{(m\mu t)^l}{l!} e^{-m\mu t} + \sum_{i=N-k+1}^{\infty} g_i \sum_{j=0}^{N-m-1} \sum_{l=0}^{j} \frac{(m\mu t)^l}{l!} e^{-m\mu t}\right] + \sum_{k=m}^{N-1} P_k \left[\sum_{i=1}^{N-k} g_i \sum_{j=0}^{i-1} \sum_{l=0}^{j+k-m} \frac{(m\mu t)^l}{l!} e^{-m\mu t} + \sum_{i=N-k+1}^{\infty} g_i \sum_{j=0}^{N-k-1} \sum_{l=0}^{j+k-m} \frac{(m\mu t)^l}{l!} e^{-m\mu t}\right]\right] \cdot \left[\sum_{k=0}^{N-1} P_k \left[\sum_{i=1}^{N-k} i g_i + \sum_{i=N-k+1}^{\infty} (N-k) g_i\right]\right]^{-1}. \quad (8)$$

The above equations are most easily understood by noting for example that the complementary waiting time distribution for the test customer in the rth position of a batch of size i, conditioned on this batch having found the system, for example, in state k where $k > m$ is

$$i g_i \bar{H}^{*(r-1+k-m)}/\text{MBS}$$

where $\bar{H}^{*(n)}$ denotes the complementary distribution function of the n-fold convolution of the service time distribution.

The conditional probability of waiting may be obtained by putting $t = 0$ in the above equations. Dividing (7) and (8) by this probability gives the waiting time distributions for the test customer conditioned both on its being accepted and on its having to wait. This eliminates the zero waiting component for directly served customers. The mean waiting times may be also obtained from (7) and (8), but it is easier to obtain them directly from Little's law using effective arrival rates [9]. If W is the mean waiting time for all customers that enter the system, and L_q is the mean queue length, then

$$W = \frac{L_q}{\lambda'}$$

where λ' is the effective arrival rate of customers, given by the following.

Strategy 1)

$$\lambda' = \lambda \sum_{k=0}^{N} P_k \sum_{i=0}^{N-k} i g_i. \quad (10)$$

Strategy 2)

$$\lambda' = \lambda \sum_{k=0}^{N} P_k \left[\sum_{i=1}^{N-k} i g_i + \sum_{i=N-k+1}^{\infty} (N-k) g_i\right]. \quad (11)$$

The queue mean length is given by

$$L_q = \sum_{k=m+1}^{N} (k-m) P_k.$$

IV. RESULTS

To demonstrate the application of the formulas of the previous section to the model of the communications controller, some numerical results are presented for the single server system. The curves are intended to be representative of those used for the buffer dimensioning performance analysis with respect to blocking and delays, and finally some throughput considerations under overload.

Dimensioning curves are shown in Figs. 2 and 3. Fig. 2 shows the effect of batch acceptance strategy, for a selection of dimensioned traffic loads. It may be seen that while acceptance strategy makes some appreciable difference for short queues, its effect diminishes with increasing queue length. Strategy 1 (accept only whole batches) always has higher customer blocking probability and, hence, lower customer throughput. The offered traffic intensity is found from

$$\rho = \frac{\lambda \cdot \text{MBS}}{m\mu}. \quad (12)$$

For large s, geometric batches and using acceptance Strategy 2, the results will be the same as in [3]. In Fig. 3, the effect of batch distribution and mean batch size is shown.

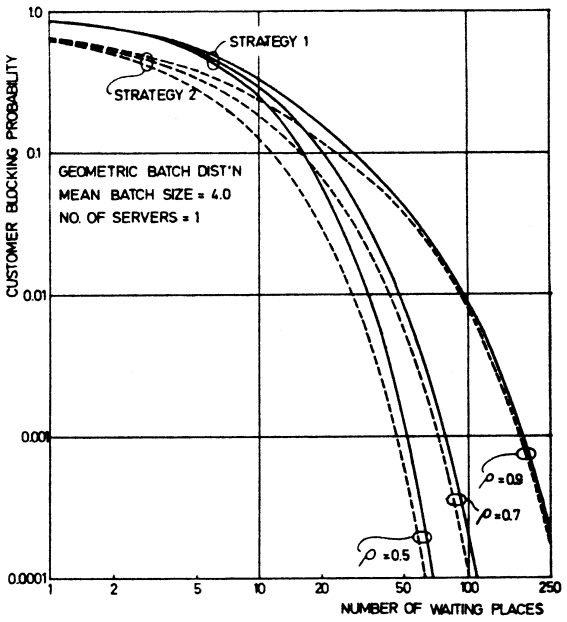


Fig. 2. Customer blocking versus queue size; parameter, acceptance strategy.

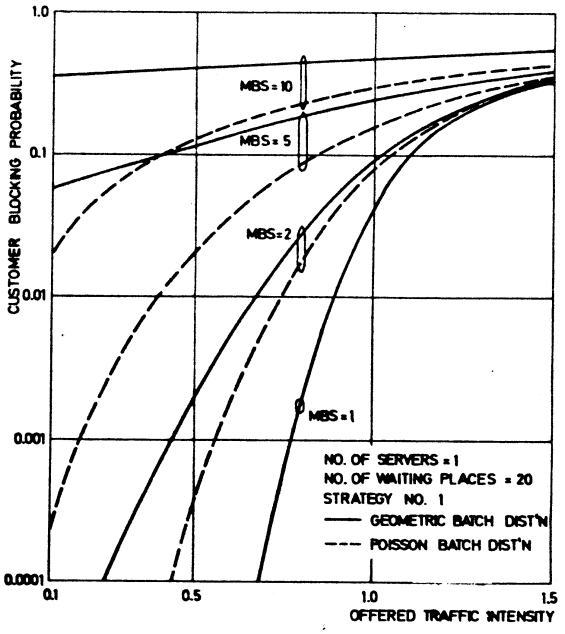


Fig. 4. Customer blocking versus offered traffic; parameter, mean batch size.

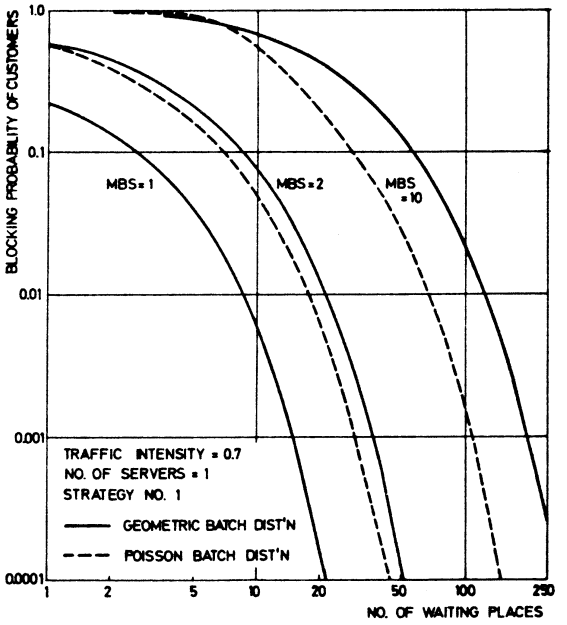


Fig. 3. Customer blocking versus queue size; parameter, mean batch size.

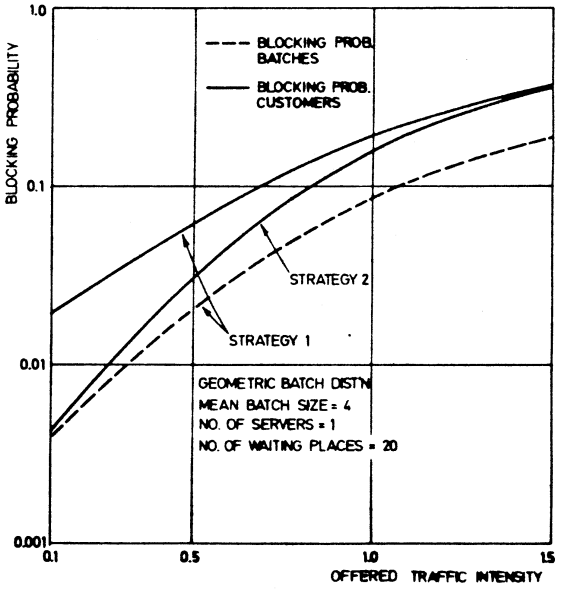


Fig. 5. Blocking versus offered traffic; parameter, acceptance strategy.

Clearly the call blocking is sensitive by orders of magnitude to the batch size parameters, so the characteristics of the input traffic must be carefully modeled. Geometrically distributed batches which have a high dispersion perform worse than the Poisson distributed batches, except in the cases of short queues and high mean batch size, which may be attributed to the relatively large number of short batches from the geometric distribution as compared to the Poisson for large mean batch size. Uniformly distributed batches were also considered, but gave results very similar to those for the Poisson.

Figs. 4–7 show performance curves for a single server system with 20 waiting places. Fig. 4 gives the effect of batch distribution on the customer blocking probability as a function of offered traffic intensity, and again the sensitivity to the batch parameters is noted. Fig. 5 has a revealing comparison between the batch and customer blocking probabilities for each acceptance strategy. For Strategy 1, the individual customer blocking probability is always higher than the batch blocking probability, since the (whole) batches rejected by this strategy will tend to be the bigger ones containing proportionally more customers. For Strategy 2, it arises that the customer and batch blocking probabilities are identical (for

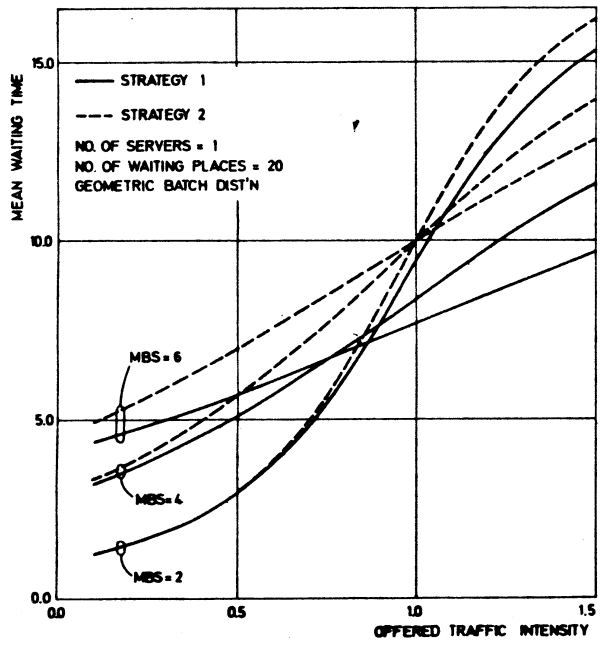


Fig. 6. Mean waiting time versus offered traffic; parameter, mean batch size.

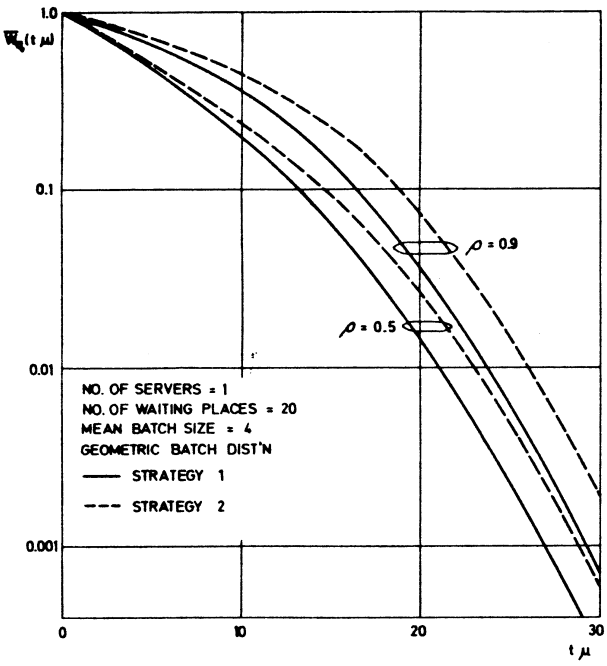


Fig. 7. Complementary waiting time distribution function.

the considered case of geometrically distributed batches). This may be seen by a reduction of (6). The relation of the Strategy 2 curve to the Strategy 1 curves may be understood by noting that at low traffic equal numbers of *batches* tend to be affected by blocking, whereas at high traffic, equal numbers of *customers*. In [3], no distinction is made between batch and customer blocking, but since in [3], the batches are assumed to be geometrically distributed and the equivalent of our Strategy 2 is used, fortunately no distinction is necessary.

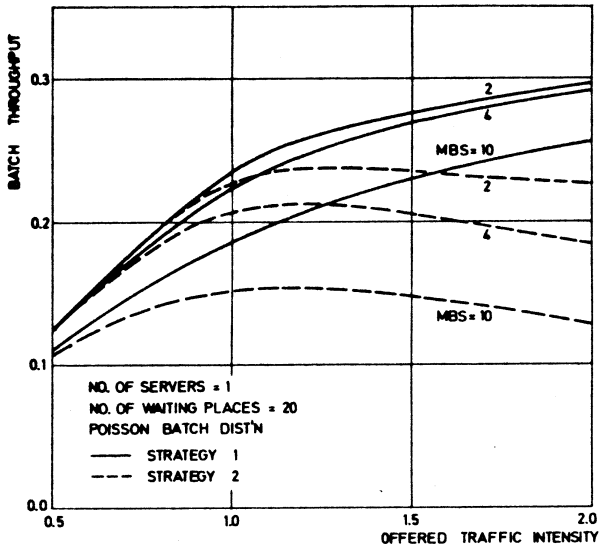


Fig. 8. Effective batch throughput versus offered traffic.

Fig. 6 shows the performance in terms of mean waiting time versus offered traffic. The mean waiting time increases with increasing offered load as expected, but an interesting crossover effect occurs as a function of the mean batch size. For both strategies, increasing mean batch size tends to increase the mean delays at low traffic loads, but actually *decreases* the mean delays at high traffic loads. At low load, there is little blocking and the increased clustering of customers for large mean batch sizes tends to increase the second component of delay, that caused by calls in the same batch. At high loads, however, the increased blocking for large batches means that relatively fewer customers are accepted for large mean batch size and the mean waiting times are relatively lower. Fig. 7 gives a few representative curves for the complementary waiting time distribution function as calculated from (7) and (8), and may be used for finding percentile delays at particular offered loads.

Finally, a brief consideration to overload performance is given by Fig. 8. Here, the throughput is measured against offered load, but a customer is only considered to be a positive part of the throughput if it belongs to a batch which was accepted in full. Hence for Strategy 2, a certain amount of the server's time is wasted by customers belonging to partly accepted batches, and leads to the situation where a saturation characteristic (decreasing throughput with increasing offered traffic) is exhibited by the system.

V. A MORE GENERAL MODEL

The strongest assumption used to produce the queueing model in Section II was that of exponentially distributed times for transmission of the (generally fixed length) packets from the communications controller to the packet switch. Some justification was provided by consideration of the effects of the communications protocol governing the packet transfer at the host-packet switch interface. In any case, the exponential server can be thought of as providing a worst case estimate for the system performance. To determine the sensitivity of

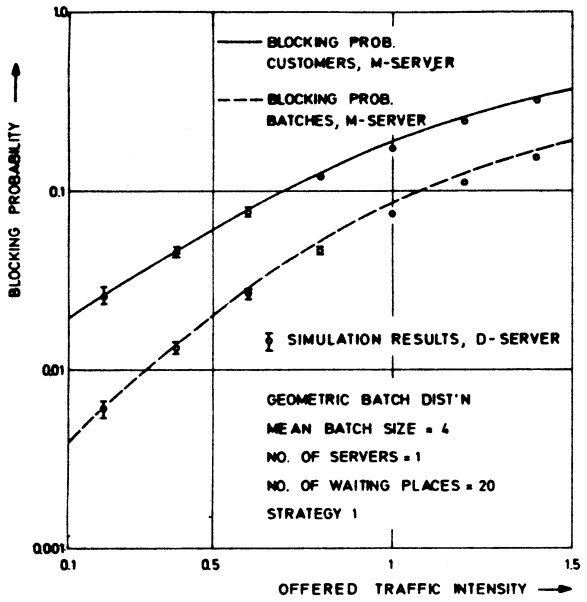


Fig. 9. Blocking versus offered traffic; parameter, service time distribution.

the batch queueing model to the service time distribution, the more general model $M^{[x]}/G/1 - s$ was investigated by means of simulation.

The results are clear in showing that all the batch queue performance characteristics are virtually *unaffected* by the service time distribution. Fig. 9 shows just a few results for the batch and customer blocking probabilities, comparing the simulation of the batch queue with fixed (deterministic) service times to the previously obtained exact results for the queue with exponential server. Each simulation result was obtained by ten part-tests each of 10 000 batches, and is given with its 95 percent confidence interval. The simulation program was validated against exact results for single-arrival queues [18] and with those for the batch queue with Markov server. Tables I and II list some further simulation results. Overall, it is clear that the queue performance is dominated by the batch size statistics, and the extreme sensitivity to service-time distribution which may be observed for single-arrival queues [18] is lost. In single-arrival queues, the fate of a test customer who arrives to find the server busy is heavily dependent on the residual service time of the customer in service, whereas for *batch* arrivals, the position of the test customer in its own batch tends to overwhelm the effect of the residual service time. The threshold of mean batch size at which the service time distribution begins to have a significant effect on performance was studied from extensive numerical results. For a queue with parameters as used in this paper, it was found that in the worst case (i.e., *D*-server) the error in customer blocking by assuming an *M*-server is of the order of 13 percent at $\rho = 0.6$ for MBS = 3, increasing to 80 percent at MBS = 2 (though the blocking is ten times less at MBS = 2).

In the actual system, the service time distribution would most probably fall between the extremes of exponential and deterministic, so whatever the true nature of this distribution, the exponential server may be assumed in order to analyze the system, with the knowledge that the results will be close to the true values, provided the mean batch size is not too small.

TABLE I
CUSTOMER BLOCKING PROBABILITY FOR $M^{[x]}/G/1$-20 (GEOMETRIC BATCHES WITH MEAN SIZE = 4)

Offered Traffic Intensity ρ	Service Time Distribution		
	M (Exact)	E_2 (Simulation)	D
0.2	0.02602	0.0247 ±0.0023	0.0263 ±0.0031
0.4	0.04641	0.0479 ±0.0037	0.0456 ±0.0026
0.6	0.08008	0.0757 ±0.0037	0.0763 ±0.0039
0.8	0.12914	0.1236 ±0.0055	0.1203 ±0.0054
1.0	0.19157	0.1819 ±0.0049	0.1743 ±0.0064
1.2	0.26140	0.2592 ±0.0078	0.2444 ±0.0053

TABLE II
MEAN WAITING TIME FOR $M^{[x]}/G/1$-20 (GEOMETRIC BATCHES WITH MEAN SIZE = 4, ONLY FOR CUSTOMERS WHO WAIT)

Offered Traffic Intensity ρ	Service Time Distribution		
	M (Exact)	E_2 (Simulation)	D
0.2	4.5383	4.439 ±0.039	4.425 ±0.044
0.4	5.4169	5.314 ±0.066	5.163 ±0.068
0.6	6.4638	6.272 ±0.063	6.138 ±0.061
0.8	7.6438	7.433 ±0.069	7.178 ±0.049
1.0	8.8905	8.705 ±0.120	8.461 ±0.099
1.2	10.1233	10.060 ±0.111	9.725 ±0.060

VI. CONCLUSION

The results in Sections IV and V are clear in showing that the most important factor affecting the system performance is the batch size statistics of the arrival process, not only the mean batch size, but also the batch size distribution. This means that in terms of the system being modeled, it is critical to measure accurately the actual size distribution of arriving messages. Generally speaking, larger message sizes and high message size variance worsen system performance.

The time distribution to transmit packets from the communications controller to the packet switch has been shown to have very little effect, and so we conclude that the initial Markovian model is sufficiently accurate to determine system performance regardless of this distribution (given that the mean batch size is not too small). This is important because

the general model $M^{[x]}/G/1 - s$ is very difficult to analyze other than by simulation. Moreover, for Strategy 1 it seems that even an imbedded Markov chain approach is not sufficient for analysis since the acceptance or rejection of batches depends not only on the sizes of arriving batches but also on the *order* in which they arrive.

The batch acceptance strategy adopted by a system also has a significant influence on performance, but the choice of strategy depends on the application of the packet-switching network. In the transferral of bulk data, it is not essential that all packets of a message stay together, so Strategy 2 could be used with its resultant higher throughput. However, for the common channel signaling packet switch of a telephone exchange (using the signaling system CCITT No. 6) it is more important that all the packets (signal units) of one message do stay together in order to preserve the integrity of the message, so Strategy 1 should be used.

Finally, the comment is made that the analysis is clearly not restricted only to the communications controller model treated here, but also to a wider class of problems involving batch arrival processes.

ACKNOWLEDGMENT

The authors would like to thank Prof. P. J. Kuehn and also H. R. Van As for helpful discussions during the course of this work.

REFERENCES

[1] W. W. Chu, "Buffer behavior for batch Poisson arrivals and single constant output," *IEEE Trans. Commun.*, vol. COM-18, pp. 613–618, 1970.

[2] W. W. Chu and L. C. Liang, "Buffer behavior for mixed input traffic and single constant output rate," *IEEE Trans. Commun.*, vol. COM-20, pp. 230–235, 1972.

[3] W. W. Chu and A. G. Konheim, "On the analysis and modeling of a class of computer communications systems," *IEEE Trans. Commun.*, vol. COM-20, pp. 645–659, 1972.

[4] P. J. Kuehn, "Multiqueue systems with nonexhaustive cyclic service," *Bell Syst. Tech. J.*, vol. 58, pp. 671–698, 1979.

[5] ——, "Analysis of switching system control structures by decomposition." in *Proc. 9th Int. Telecommun. Conf.*, Spain, 1979.

[6] M. Langenbach-Belz, "Two-stage queueing system with sampled parallel input queues," in *Proc. 7th Int. Telecommun. Conf.*, Stockholm, Sweden, 1973.

[7] H. G. Schwaertzel, "Serving strategies of batch arrivals in common-control switching systems," in *Proc. 7th Int. Telecommun. Conf.*, Stockholm, Sweden, 1973.

[8] U. N. Bhat, "Imbedded Markov chain analysis of single server bulk queues," *J. Aust. Math. Soc.*, vol. 4, pp. 244–263, 1964.

[9] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. New York: Wiley, 1974.

[10] I. Kabak, "Blocking and delays in $M^{(n)}/M/c$ bulk queueing systems," *Oper. Res.*, vol. 16, pp. 830–840, 1968.

[11] A. Kuczura, "Batch input to a multiserver queue with constant service times," *Bell Syst. Tech. J.*, vol. 52, pp. 83–99, 1973.

[12] R. G. Miller, "A contribution to the theory of bulk queues," *J. Roy. Stat. Soc. B.*, vol. 21, pp. 320–337, 1959.

[13] T. P. Bagchi and J. G. C. Templeton, "Finite waiting space bulk queueing systems," *J. Eng. Math.*, vol. 7, pp. 313–317, 1973.

[14] S. I. Rosenlund, "Busy periods in time-dependent $M/G/1$ queues," *Adv. Appl. Prob.*, vol. 8, pp. 195–208, 1976.

[15] K. Pawlikowski, "Message waiting time in a packet switching system," *J. Ass. Comput. Mach.*, vol. 27, pp. 30–41, 1980.

[16] R. J. Cypser, *Communications Architecture for Distributed Systems*. Reading, MA: Addison-Wesley, 1978.

[17] P. J. Burke, "Delays in single-server queues with batch input," *Bell Syst. Tech. J.*, vol. 54, pp. 830–833, 1975.

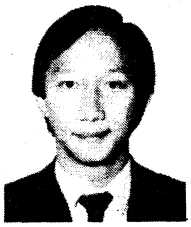
[18] P. J. Kuehn, "Tables on delay systems," Inst. Switching and Data Technics, Univ. Stuttgart, Stuttgart, Germany, 1976.

★

David R. Manfield (S'78–M'80) was born in Brisbane, Qld., Australia, in 1955. He received the B.E. (1st Cl. Hons.) degree from the University of Queensland, Brisbane, in 1975 and the Ph.D. degree in 1980, both in electrical engineering.

From April 1980 until February 1981 he was working in the Department of Communications, University of Siegen, Siegen, West Germany under a postdoctoral fellowship from the Alexander von Humboldt Foundation, on topics concerned with queueing analysis of switching systems. He is now working with Bell-Northern Research, Ottawa, Ont., Canada, where he is involved in traffic engineering studies for SPC switching systems.

★



P. Tran-Gia (M'80) was born in Vietnam. He received the M.S. degree (Dipl.-Ing.) in electrical engineering from Stuttgart University, Stuttgart, West Germany, in 1977.

In 1977, he joined Standard Elektrik Lorenz (ITT), Stuttgart, where he was working in software development of digital switching systems. Since 1979, he has been Assistant Professor at the Department of Communications, University of Siegen, Siegen, West Germany. His current research activities are in the field of queueing theory and its application in performance analysis for telecommunication systems.