AN APPROXIMATE MODEL AND ANALYSIS FOR A QUEUE WITH HIGHLY CLUSTERED ARRIVALS

Harry Rudin, Jr.
IBM Research Laboratory Zurich
Rüschlikon, Switzerland

ABSTRACT

There is a growing number of processes wanting analysis in which events arrive in a fashion more clustered in time than would be predicted by Poisson statistics. The purpose of this paper is to present an analytic solution for a model which, while approximate, nevertheless retains the essential characteristics of the burst arrival spread over a time interval of nonnegligible length.

Specifically, a single server queue of unlimited length is considered. The service times are exponentially distributed and the service discipline is first-in-first-out. The arrival process is composed of two Poisson processes: the one with the higher arrival rate is effective when the queue (including server) is nonempty; the one with the lower arrival rate is effective when the queue is empty. For this model the average delay for waiting time only is simply through the ratio of the two arrival rates just described.

The analytic solution for the model is compared with simulations. As expected the analytic results are somewhat pessimistic with respect to expected delays but nevertheless properly characterize the behavior of the queue with clustered arrivals.

While the problem of clustered arrivals was the focus of this work, the results are also applicable to the situation where some arrivals balk at joining a queue. This is handled by interchanging the two arrival rates discussed above.

1. INTRODUCTION

There is a growing number of processes wanting analysis in which events arrive in a fashion more clustered in time than would be predicted by Poisson arrival statistics. Examples from the telecommunication area are: 1) job stream arrival within a data-processing system; 2) scanning sensor output in a raster-scanned facsimile system; and 3) arrival of packets in a message-switching telecommunication system.

When these highly clustered events arrive at a queueing point the result is much more congestion than would be predicted by an analysis assuming Poisson arrival statistics. This manifests itself as substantially longer delays and in the case of finite queueing capacity - higher probability of queue overflow.

While the growing importance of the problem is a recent development, the problem is not a new one. The occurrence of clustered arrivals introduces dependencies or correlations into the analysis of queueing systems. A general survey of queueing systems with such dependence is given by Bhat. Marlin2 has a list of references specifically for dependence in the arrival process.

One means of handling the clustered arrival phenomenon is through the assumption that many customers arrive simultaneously in a single burst. Such analyses have been made by Restrepo3 for Erlang service time distributions and by Hawkes4 for more general service time distributions and including priorities. Burst arrivals in the case of finite length queues (with overflow) have been examined by Chu5 for a "compound Poisson process" and by Rudin6 for binomial arrivals. References 3 through 6, however, result in one-dimensional distribution functions for event arrival. While these approaches are useful, they cannot accurately account for the Markov nature of the clustered arrival process wherein the arrival of one event markedly increases the probability of the arrival of another event in the ensuing time interval. It is specifically this kind of process which we are trying to model.

A well-documented case of clustered arrivals is the generation of characters and messages at typewriter terminals in a conversational teleprocessing system. For such sources Gordon, Wetzler and Pilc5 have carried out some computer simulations of a statistical multiplexer and have carefully modeled the input process as being composed of a number of these sources. Their study has also underscored the importance of burstiness or clusters in the input traffic. Chaudhry9 provides an analysis for a discrete-time queue with exponential service time and where there is correlation between arrivals at two consecutive discrete "time marks".

In the present paper a simple model for the clustered arrival process is suggested. An exact analytic solution is found for this approximate model in the hope of providing a simple relation of engineering significance.

2. THE MODEL AND ITS ANALYSIS

The model has an input process characterized by two states of relative activity. The input process is shown in Fig. 1. When the bursty input process is in the active state S1 the arrival rate is \( \lambda_1 \); when the input process is in the in-
active state \( S_0 \), the arrival rate is \( \lambda_0 \). In an actual situation whether a source is in the inactive state or in the active state depends on characteristics of the source alone; i.e., it is independent of the queue as will be simulated. The approximation which is made in the analysis here is that if the queue including server is empty, the arrival process is in the relatively inactive state; if the queue is nonempty (again including the server) the arrival process is in the active state \( SI \).

In making this assumption, the worst kind of interdependence is introduced in the sense that the congestion is maximized. It is therefore anticipated that the amount of congestion or delay predicted by this approximate model is a worst case. It will be seen that this expectation is corroborated by simulation of a system where the bursty source is independent of the queue state.

A parameter in the analysis is the utilization \( \rho \), the fraction of time which the server is occupied. \( \rho \) is also the probability of finding the system in state \( SI \). Since the two states enumerated are mutually exclusive, the probability of the system being in state \( S_0 \) is \( 1 - \rho \). This is noted in Fig. 1.

The average arrival rate is then given by

\[
\lambda = \text{Probability (queue empty)} 	imes \lambda_0 + \text{Probability (queue nonempty)} 	imes \lambda_1 = (1 - \rho) \times \lambda_0 + \rho \times \lambda_1. \tag{1}
\]

It is also assumed that the service times are distributed negative exponentially with average length \( T = 1/\mu \). The analysis of the steady state proceeds simply by writing the equations of detailed balance, as for example in Morse. \( ^{10} \) \( P_i \) is the probability of the queue (including server) being in state \( i \),

\[
P_0 \lambda_0 = P_1 \mu, \tag{2}
\]

\[
P_0 \lambda_0 + P_2 \mu = P_1 (\lambda_1 + \mu), \tag{3}
\]

\[
P_{n-1} \lambda_1 + P_{n+1} \mu = P_n (\lambda_1 + \mu), \quad n \geq 2. \tag{4}
\]

Equations (2) - (4) can be rewritten

\[
P_1 = \frac{\lambda_0}{\mu} P_0 \tag{5}
\]

\[
P_2 = \frac{\lambda_1}{\mu} P_1 - \frac{\lambda_0}{\mu} P_0 \tag{6}
\]

\[
P_{n+1} = P_n + \frac{\lambda_1}{\mu} P_n - \frac{\lambda_1}{\mu} P_{n-1}, \quad n \geq 2. \tag{7}
\]

Solving successively and knowing \( P_0 = (1 - \rho) \) as above,

\[
P_0 = (1 - \rho) \tag{8}
\]

\[
P_1 = \frac{\lambda_0}{\mu} P_0 = \frac{\lambda_0}{\mu} (1 - \rho) \tag{9}
\]

\[
P_n = \frac{\lambda_1}{\mu} P_{n-1} = \left( \frac{\lambda_1}{\mu} ight)^{n-1} \left( \frac{\lambda_0}{\mu} \right) (1 - \rho), \quad n \geq 2. \tag{10}
\]

These last equations can be simplified using (1) and the relation \( \rho = \lambda/\mu \) to yield

\[
P_0 = (1 - \rho) \tag{11}
\]

\[
P_n = \left( \frac{\lambda_1}{\mu} \right)^{n-1} \left( 1 - \frac{\lambda_1}{\mu} \right) \rho, \quad n \geq 1. \tag{12}
\]

In Fig. 2 the probabilities of being in the various states are plotted as a function of those states; all curves are drawn for a utilization of \( \rho = 0.75 \). The two solid curves are calculated from equations (11) and (12) and have meaning only for integer values of the abscissa. The curve for \( \lambda_1/\lambda_0 = 1 \) degenerates to the standard M/M/1 solution. The effect of a \( \lambda_1/\lambda_0 \) ratio greater than unity is clear.
where $T_s$ is the average service time, $1/\mu$. The waiting time thus varies from the conventional $M/M/1$ case by the ratio of the two arrival rates, $\lambda_4/\lambda_0$. Further the average queueing time (including service time) is

$$T_q = \left\{ 1 + \frac{\rho \cdot \lambda_1}{1-\rho} \right\} \times T_s \quad \text{(15)}$$

The result is a simple analytic (but approximate) description of the complicated clustered arrival process.

3. CRITIQUE: COMPARISON WITH SIMULATION RESULTS

Since the model is an approximate one its behavior must be compared with the performance of a more realistic model, i.e., one without interdependence between the arrival process and queue state. This comparison has been carried out by constructing a model in the GPSS language.

In the simulated model the same convention is observed as is indicated in Fig. 1, i.e., the probability of the arrival process being in state $S_0$ is $(1-\rho)$ and the probability of being in state $S_1$ is $\rho$. In the analytic model the transitions between the two states are determined by the emptiness or nonemptiness of the queue (again, including server). In the simulation model transitions between the two states are independent of the queue state and in fact these transitions are controlled by an additional Poisson process. It is only at the arrival times of the events in this fictitious process that transitions between the states $S_1$ and $S_2$ are allowed. The transition probabilities themselves are determined in accordance with the two state probabilities, $\rho$ and $(1-\rho)$. In summary the arrival process for the queue we are simulating consists of periods of relative activity and inactivity controlled by an external Poisson process with known average time between event arrivals.

Figure 3 shows the average waiting time as a function of the average time between allowable transitions as controlled by the additional Poisson process just described. Both the scale for average waiting time and the scale for average time between allowable transitions are normalized so that $T_s = 1$. The upper portion of the figure is drawn for a utilization of 25% and a ratio of $\lambda_4/\lambda_0$ of 10. For these values equation 14 predicts an average waiting time of 3.33 $T_s$. This value is shown as a dashed horizontal line at the top; the standard $M/M/1$ analysis without bursts yields an average waiting time of 0.33 $T_s$, shown as a dashed line at the bottom of the figure. The crosses indicate the results of the simulation. It can be seen that when the average time interval between allowable transitions is comparable to the service time, the average delay in the system is approximately equal to the non-burst analysis. As the time between allowable transitions increases, the actual average waiting times increase and approach but never reach the delay predicted by the analysis of the approximate model.

This is not true for the case shown at the bottom of the figure where the predicted values for bursty arrivals are obtained for long duration times in the two states. Parameters for the lower portion of the figure are utilization of 75% and ratio $\lambda_4/\lambda_0 = 4$. The queue state probabilities have been measured in a simulation where the average time between allowable transitions was 100 times the average service time and are compared with the analytic results for the approximate model in Fig. 2. The results are pleasing.

In all that has been said here so far, the implicit assumption has been that we have modeled the input process as essentially a single source. In the teleprocessing environment it is likely that the input to a queue derives not from a single source but from a number of sources. Figure 4 is an attempt to shed some light on this point. Here average total delay is plotted as a function of the ratio $\lambda_4/\lambda_0$; the utilization is 75%. The theoretical result from equation 15 is shown as a dashed line near the top of the figure. Also shown are results for the simulation of a single source (already discussed), for two independent sources, and for four independent sources. In all cases a line is drawn fitted to the points obtained by simulation and the average time between allowable transitions was 100 times the average service time. As expected, as the number of bursty sources increases, the effect of burstiness in the conglomerate input process decreases.

In concluding this section we observe that, in general, the analysis of the approximate model is pessimistic. Real cases approach the analytic results when the input process can be modeled as a single process and when this single process remains in its active and inactive states for relatively long periods of time compared with the average service time.
Average Total Delay

![Diagram showing analytic and simulation results for different source configurations]

Fig. 4. Analytic and simulation results showing the effects of several sources: total delay as a function of $\lambda_1/\lambda_0$.

4. EXTENSION TO BAWKING

The emphasis of this analysis has been on the problem of clustered arrivals. There exists, however, the antithesis of this arrival process. This appears in the circumstance when customers bawk at joining a nonempty queue and the analysis is concerned only with the waiting time of those customers who actually do join the queue. One can apply the model developed here to this case as well; in this case the active state coincides with the event of an empty queue and the inactive state with the event of a nonempty queue. Very small ratios of $\lambda_1/\lambda_0$ correspond to extremely short waiting times in the case of bawking. Much work has been done on the bawking problem; several references may be found in Bhat. ¹

5. ACKNOWLEDGMENT

The author acknowledges with thanks numerous, helpful discussions with his colleagues, particularly H.R. Mueller.

REFERENCES


