MODELING PACKET ARRIVALS FROM ASYNCHRONOUS INPUT LINES

V. RAMASWAMI and Guy LATOUCHE

Bell Communications Research, Morristown, NJ 07960, USA, and Université Libre de Bruxelles, Boulevard du Triomphe, C.P. 212, 1050 Bruxelles, BELGIUM

We develop a unified stochastic model to describe the arrival of packets to a packet network with asynchronous input lines. Substantial generality is attained in describing the alternating sequence of talk spurts and silences in the individual call scenario. For the overall packet stream of a call type, we characterize the stationary version and derive many interesting quantities thereof. Besides laying the groundwork for developing a set of useful tools for performance analysis, our model provides a systematic method to build a load box to a packet switch.

1. INTRODUCTION

The packet stream arriving to a broadband packet switch is the superposition of the packets of individual calls, and usually there are different call types to consider. Driven by the randomness of the call mix and by the fluctuations in the packet generation by individual calls, the overall packet stream is very complicated. It is not clear whether and when simple models such as the Poisson or the interrupted Poisson process will suffice. The arrivals can be more bursty than in a Poisson process due to the simultaneous presence of many calls of possibly different types and also due to the inherent burstiness of some call types. Yet the process is not so bursty as a process with batch arrivals, for, the successive packets of a call are spaced out in time. Also, correlations in the overall packet arrival process appear, in general, to be nonnegligible, and this may rule out the use of renewal processes and in particular the interrupted Poisson process traditionally used to model bursty streams. We believe that there is indeed a genuine need for developing a new set of tools for describing such processes.

Our goal here is to develop a point process model that takes into account in a general manner the statistical fluctuations in the call set up times as well as those in the scenario of packet generation by individual calls. Needless to say, such a process is too general to be tractable when used as an input to queueing models. Thus, some approximations will indeed be necessary for future performance work. However, it is hoped that the model developed here can help: (a) to identify the essential features of the packet stream that have a significant effect on performance so that approximations that capture such effects can be identified, (b) to examine, in particular, the efficacy of approximating the process by a Poisson or an interrupted Poisson process, (c) to identify the particular call scenarios that result in undesirable performance, and (d) to provide a systematic method to build a "load box".

A particularly useful aspect of our model is the unified manner in which different call types are described. With each call type we associate a Poisson process to describe call set up times (extension of our results to more general Markovian point processes appears possible), and assume that each call within a call type evolves according to a discrete time Markov chain which describes the scenario of its packet generation process. Different call types are modeled by modifying appropriately the parameters of these processes, and this allows for retaining a common structure. Let us start our discussion with the Markov chain model for a single call.

2. ANATOMY OF A CALL

The description to follow relates to calls of a common type. Calls of different types are characterized by different parameters.

4.1A.5.1
Each call has a random holding time comprised of an alternating sequence of 'talk spurts' (active periods) and 'silences' (inactive periods). During talk spurts, a call generates packets at equidistant time points separated by the amount of time it takes to gather a packet. For example, a voice call on a 64 Kb/s line takes 16ms to make a packet of 128 bytes.

In modeling this complex scheme, we wish to achieve the greatest generality in the distributions of the holding time, the number and duration of talk spurts, and the duration of silences. Clearly, one needs a discrete time model, and we proceed as follows.

For each call type, we choose a number $\delta > 0$ as a parameter discretizing time; $\delta$ is the amount of time to gather a packet. A call arriving at time $t$ will terminate at time $t + X \delta$, where $X$ is a positive integer valued random variable having a discrete phase type distribution (PH distribution) $PH(\alpha, T)$. Due to their importance here, we present a brief summary of the essential ideas concerning PH distributions; refer to [3] for a detailed discussion.

**Phase Type Distribution $PH(\alpha, T)$:** Consider a Markov chain (MC) on \{1, ..., $m+1$\} with initial probability vector $(\alpha_1, ..., \alpha_m, 1 - \sum_{i=1}^{m} \alpha_i)$, and transition matrix $P = \begin{bmatrix} T & t^o \\ 0 & I \end{bmatrix}$, where $T$ is an $m \times m$ substochastic matrix, $I$ is a column vector of 1's, and $t^o = 1 - T I$. Assume that states 1, ..., $m$ are transient, i.e., $(I - T)^{-1}$ exists. The r.v. $X$ is said to follow the distribution $PH(\alpha, T)$, if $X$ has the same distribution as the number of transitions in the above MC until it gets absorbed in state $m+1$. Thus, for each call, entry into the state $m+1$ by its associated MC corresponds to the termination of that call. It is trivial to verify that $P(X = n) = \alpha T^{n-1} t^o, n \geq 1$ and that the $k$-th factorial moment $\mu^{(k)} = E[X(X-1)...(X-k+1)] = k! \alpha T^{k-1}(I - T)^{-1} I, k \geq 1$. An attractive feature of PH distributions is that they lead to easily implementable matrix formulae for most quantities of interest.

PH distributions include as special cases many commonly used distributions such as the geometric, negative binomial, their convolutions and mixtures. Further, any probability distribution on the nonnegative integers with a finite support is a PH distribution. From this it also follows that the class of PH distributions is dense in the class of all probability distributions on the nonnegative integers. Thus, our model allows the greatest flexibility in modeling the holding times of a call. That, however, is not the only advantage as the following discussion will show.

**Detailed Call Scenario:** Having associated an absorbing MC with each call, we further refine our model as follows. We assume the set of transient states to be partitioned into two disjoint subsets $A$ and $S$ ($S$ may be empty) such that: (a) if the MC is in $A$, then the call is active and will submit a packet at the next epoch of transition of the MC, (b) if it is in $S$, then after a sojourn of $\delta$ units of time in that state, it may make a transition but will not submit a packet to the system at the epoch of transition. Thus, $A$ is the set of active states, and $S$ is the set of inactive states.

The overall arrival stream of packets of a call type results from the superposition of as many independent MCs as there are calls of that type in progress. Indeed, if at time $t$ there are $k$ calls in progress of which $k_A$ have their associated MCs in $A$, then exactly $k_A$ packets will be submitted to the system during $(t, t+\delta)$. By time $t+\delta$, all $k$ MCs would have undergone one transition (some may have entered the absorbing state, i.e., the corresponding calls would have terminated), and possibly new calls will have arrived.

Before we proceed with the mathematical analysis of the general arrival stream described above, let us illustrate its versatility for modeling purposes through some simple examples. In each of these, we specify the MC by its flow diagram.

**Example 2.1:** Figure 1 corresponds to the case where each call generates a geometrically distributed number of packets and terminates. There are no silent periods here.

**Example 2.2:** Figure 2 corresponds to the case where each call comprises of an alternating sequence of talk spurts and silent periods each of which is geometrically distributed.
By suitably defining $A$ and $B$ and the transition rules, one can generate different distributions for active and silent periods. We illustrate this with a very general example.

Example 2.3: In Figure 3, the alternating talk spurts and silent periods have general distributions given by the densities $\{\alpha_i\}$ and $\{\beta_i\}$ respectively.

Example 2.4: In Figure 4, each talk spurt contains 1 or 2 packets with respective probabilities $\alpha_1, 1-\alpha_1$. If a talk spurt is followed by a silent period, then the latter has the same duration as the immediately preceding active period.
3. SOME DISTRIBUTIONS RELATED TO A CALL

In conformance with the partitioning of the states, we also partition \( \alpha, T \) and \( T^o \) as \( \alpha = (\alpha_A, \alpha_S) \).

\[
T = \begin{bmatrix}
T_{AA} & T_{AS} \\
T_{SA} & T_{SS}
\end{bmatrix}, \quad \text{and} \quad T^o = \begin{bmatrix}
t_1^o \\
t_2^o
\end{bmatrix}.
\]

We assume throughout that \( \alpha_s = 0 \) (i.e., calls start with a talk spurt), and that \( t_s^o = 0 \) (i.e., calls end with a talk spurt). Under these assumptions motivated by practical considerations, we present a number of results which are easily modified when these conditions are not met. The following theorems show that not only the holding time but also many other random variables related to a call have PH-distributions whose parameters are easily determined. This opens up interesting possibilities for developing refined models which match many interesting characteristics besides those of only the call holding time. For proofs, we refer the reader to [4].

**Theorem 3.1:** The holding time of a call is \( X \delta \), where \( X = \text{PH}(\alpha, T) \).

**Theorem 3.2:** \( K \), the total number of packets submitted by a call, is such that \( K = \text{PH}(\alpha^*, T^*) \), where \( \alpha^* = \alpha_A \), and \( T^* = T_{AA} + T_{AS}(1 - T_{SS})^{-1}T_{SA} \). The total active time of the call is \( K \delta \).

**Theorem 3.3:** The total duration of silent periods in the holding time of a call is distributed as \( Z \delta \), where \( Z = \text{PH}(\beta, U) \), where \( \beta = \alpha_A (1 - T_{AA})^{-1}T_{AS} \), and \( U = T_{SS} + T_{SA}(1 - T_{AA})^{-1}T_{AS} \).

**Theorem 3.4:** The total number \( M \) of talk spurts during the holding time of a call has distribution \( \text{PH}(\alpha, T) \), where \( \alpha = \alpha_A \), and \( T = (1 - T_{AA})^{-1}T_{AS}(1 - T_{SS})^{-1}T_{SA} \). The total number of silent periods is \( M - 1 \) and has PH-distribution \( \text{PH}(\alpha^*_T, T) \).

We have noted that in general successive talk spurts and silent periods may be correlated. A general nsr for their independence is not easy to obtain. In the particularly interesting situation where each of the subsets \( A \) and \( S \) has only one exit state, clearly the successive talk spurts and silent periods are independently distributed. Also, in the further interesting special case where the exit state \( i \in S \) is such that \( T_{ij} = c \alpha_j \) for all \( j \in A \) for some \( c > 0 \), it is easy to see that successive talk spurts and silent periods form an alternating renewal sequence. Even within this highly restricted class of models, one can obtain a wide variety of scenarios by changing the distributions of the silent and active periods.
Having illustrated the versatility of the general model of the call scenario, let us now turn to the superposition process.

4. THE STATIONARY PROCESS OF PACKET ARRIVALS

To completely characterize the overall packet stream due to a call type, we must specify at each time $t \geq 0$: $n(t)$, the number of calls in progress at time $t$; for each call $i$ in progress, the length $0 < x_i(t) \leq \delta$ of the remaining interval of time until its associated MC will undergo a transition; and finally for each call $i$, the last state $s_i(t)$ in which its MC entered before time $t$. To determine the joint distribution of these quantities, we note that calls in progress may be described by an $M/G/\infty$ queue whose service time is the holding time of a call.

From known results for $M/G/\infty$ queues, see [6], p161, the stationary distribution of $n(t)$ is Poisson with parameter $\theta$, where $\theta = \delta \alpha(l-T)^{-1}$ is the mean holding time. Further, the asymptotic remaining holding times, given the number of calls in progress, are iid random variables with common distribution $\theta^{-1} \int [1-H(u)] du$, where $H(\cdot)$ is the holding time distribution.

By a direct computation using

$$1 - H(t) = P[X \geq (k+1)] = \alpha T^k 1, \quad \text{for } k = 0, 1, 2, \ldots \text{ and } k \delta \leq t < (k+1)\delta,$$

one sees that the density function of the remaining holding time of a call in progress is given by

$$g(t) = [(\delta \alpha(l-T)^{-1}) \alpha T^k 1 = \theta^{-1} \alpha T^k t^k, \quad k \geq 0, \quad k \delta \leq t < (k+1)\delta,$$

where $\alpha = [\alpha(l-T)^{-1}]$. From this, the following theorem is immediately obtained; for details of the proof see [4].

Theorem 4.1: The stationary process of packet arrivals (of a call type with arrival rate $\lambda$ and associated MC governed by $PH(\alpha, T)$) is stochastically equivalent to the following:

a) The number of calls in progress at time 0 has a Poisson distribution with parameter $\theta$, where $\theta$ is the mean holding time of a call.

b) Each call $i$ in progress at time 0 chooses an initial state and a delay $\tau_i$ independently of each other such that the initial state has distribution $\pi$ and the delay has the uniform distribution $U(0, \delta)$.

c) For call $i$, the MC associated with that call will undergo its first transition at the epoch $\tau_i$ chosen as above, and the call will submit a packet at $\tau_i$ iff its initial state $= A$. Having chosen the initial state and the epoch of transition $\tau_i$, the associated MC makes transitions at $\tau_i + k\delta$, $k \geq 0$ until it gets absorbed in state $m+1$ (call termination) as outlined in Section 2.

d) New calls arrive according to a Poisson process with parameter $\lambda$. The initial phase for each new call has distribution $\pi$, and the delay is $\delta$ a.s. From the epoch of arrival, the packet arrivals due to that call are governed by a MC as described in Section 2.

5. THE INDEX OF DISPERSION CURVE

Based on the characterization in Theorem 4.1 of the stationary version of the packet stream of a call type and using standard computations of the type used in Markov renewal theory, we have derived formulae for the first two moments of the quantity $N(t)$, the number of packets arriving in the interval $[0, t]$. While the formula for the second moment is quite complicated, nevertheless it is eminently computable. For reasons of brevity, we must refer the reader to [4] for the applicable formulae and algorithms. Using these algorithms, we computed the Index of Dispersion Curves for a number of examples, and we restrict our discussion to some salient features brought out by their examination. As is well-known [1], the index of dispersion curve is obtained by plotting as a function of $t$ the quantity $I(t) = Var[N(t)]/E[N(t)]$. This quantity is identically 1 for the Poisson process, and is in general a measure of the burstiness of the process. An important reason for examining this curve is that it provides information about the correlation structure in the process; this is due to the fact that $Cov[N(t), N(2t) - N(t)] = 0.5 Var[N(2t)] - Var[N(t)]$. 

4.1A.5.5
In Figure 5 are plotted the index of dispersion curves for two examples. In each of these examples, there are no silent periods. The number of packets submitted by a call has respectively a geometric and a negative binomial distribution with the same mean. These examples were generated assuming a mean call holding time of 2 minutes, $\delta = 16 ms$, and an overall packet arrival rate that corresponds to a load $\rho = 0.6$ on a 150 Mb/s output line. Even a cursory examination shows that the plotted curves are quite different from what one expects under a Poisson model; for the latter we noted that $I(t) = 1$. Also, note that these curves take a long time to reach their asymptotes; we shall discuss this point soon.

Our model permits us to examine different call scenarios. This is illustrated in Figure 6. In all its examples, the total number of packets submitted has the same geometric distribution. However, we consider the effect of silences interspersed between talk bursts. By varying the mean times of talk spurts and silences, three scenarios are generated to compare against the case with no silence. Although the asymptotes for all the curves can be shown to be the same, it is interesting to see the substantial differences in the index of dispersion curves depending on the call scenario.

That the overall process of packets should exhibit substantial correlations appears to be intuitive. Shorter interarrival times are usually the effect of a large number of calls being in progress, and in this manner the random environment induces burstiness in the arrival stream. In general, one finds it difficult to accept that a Poisson process or even a renewal process could model the arrival process adequately. Conflicting this are, however, the conclusions of a recent study [5] by Sriram and Whitt. Assuming a constant number of calls in the system, they argue that under light to moderate loads, the arrival process can be well approximated by a Poisson process. By
careful simulation experiments and theoretical reasoning it is also shown that the long term covariances exert significant influence in higher traffic intensities rendering the Poisson model inadequate as an approximation.

A consequence of Theorem 4.1 is the following. Given \( n \) calls at time \( t \) that are in the set \( A \) (i.e. in a talk spurt), the arrival epochs in the interval \([t, t+\delta]\) form an ordered sample of size \( n \) from the uniform distribution \( U(0, \delta) \). From this one can show that as \( n, \delta \to \infty \) in such a way that \( n/\delta = \gamma \), the joint distribution of the first (fixed) \( k \) interarrival times converges to that of a product of exponential distributions. This is in conformity with the results of Sriram and Whitt. If the \( k \)-th arrival epoch is much smaller than \( \delta \) and if there are a large number of calls, then we may consider the first \( k \) interarrival times as those from a Poisson process. However, this creates a dilemma. For the limit theorem to hold, we need a large \( n \) and short interarrival times compared to \( \delta \), but would that not put us automatically in a heavy traffic situation where correlations matter? Another difficulty is that if \( n \) is large, can we ignore the fact calls can terminate - i.e. can we pretend that we have a constant number of calls in progress? As rightly pointed out in [5], the important question is: how large is large? We do not consider this to be an easy problem to resolve.

We have also approached the approximation issue from the point of view taken in [2], where a Markov modulated process matching the index of dispersion curve in its entire range is taken as an approximation. Based on our moment formulae, we have been able to prove that \( I(t) \to E(K^2)/E(K) \) as \( t \to \infty \), where \( K \) is the total number of packets delivered by a call during its entire holding time. This result shows, among other things, that the limiting value of \( I(t) \) provides little information on the burstiness of the process; that value does not depend on the spacings between packets of an individual call. Combined with the fact that the curves shown in Figures 5 and 6 take a long time to reach their asymptotes, it appears that one should not attach too much importance to this limit. Then what is the interval over which one must match the index of dispersion curve?

These questions do not appear to be resolvable by theoretical arguments only. The limit theorems have to be refined to yield approximation theorems, but such results are not easy to obtain. One promising approach is an experimental/empirical exploration based on system measurements obtained by offering an artificial load to an existing switch. The model presented here appears to offer a systematic method to build such a load box, and therein may lie its importance for future research.

REFERENCES

[4] Ramaswami, V and Latouche, G: A unified stochastic model for the arrival of packets to a broadband switch, Bell Communications Research internal memorandum.