A Fair Statistical Multiplexer For An Integrated Services Packet Network

Ioannis Stavrakakis
EE/CS Department
University of Vermont
Burlington, Vermont 05405

Abstract

In this paper a statistical multiplexer with N input lines is analyzed under a general packet arrival model. The model is suitable for the unified description of the packet processes generated by different sources of information in an Integrated Services Packet Network (ISPN). To introduce a kind of fairness in the service procedure, it is assumed that the first packet of a message (or one of the packets arriving at the same slot) is given priority by entering a preemptive priority queue while the rest of the packets enter a second (low priority) queue. This service policy favors single packets or, in general, short messages over long ones. In an Integrated Services Network, where different kind of messages are accommodated, it may be necessary to adopt such a service policy to avoid long delays of single packets (or short messages).

The mean delay of the two categories of packets is calculated together with the mean buffer occupancy in each of the two queues.

I. Introduction

Integrated Services Packet Networks (ISPN's) should not be seen as a simple evolution of Data networks (DN's) which have been developed over the last two decades. The significant differences among the sources of information involved in an ISPN, regarding, for instance, the packet generation processes and the packet delivery requirements, create a significantly different environment from that in DN's.

Although the unit of information is a fixed size packet for all potential users of the system, to facilitate the integrating operation of an ISPN, the characteristics of the various packet processes of interest can be dramatically different from those present in a traditional DN. Poisson, Bernoulli, or general i.i.d. processes, widely incorporated in the analysis of DN's, are rather inappropriate for the description of the packet processes in an ISPN. For instance, packetized voice traffic can be modeled as blocks of packets arriving over consecutive time slots with geometrically distributed length (talkspurt) followed by periods of silence with geometrically distributed length. Other kinds of packetized information (such as long files, video traffic, etc) may be described as blocks of packets of length following a general distribution. The output of a computer over a slot may contain more than one packets of information; fast transmission lines may also deliver more than one packets per slot. The packet traffic generated by a concentrator/transmitter and being delivered through a slotted line is constant (one packet per slot), whenever its buffer is non-empty and it is zero otherwise. Packet traffics generated by various sources in an ISPN or by network components in both an ISPN or a DN cannot be described with the memoryless models mentioned before.

In a discrete time slotted network, the packet traffics for the cases described above (among other) can be appropriately described by a Markov Modulated Generalized Bernoulli Process (MMGBP). That is, it is assumed that the source of information (i.e. network component or user) visits M states of an underlying Markov chain. Given the current state, the number of packets generated follows a general distribution. Clearly this packet process is a non-i.i.d. one. It is easy to establish that the cases of packet traffics described before may be described (or approximated) by a MMGBP. For instance, the packetized voice traffic is a MMGBP with two states, "talkspurt" and "silence". The probability that the voice source generates one packet when in state "talkspurt" is one; the probability that it generates zero packets when in state "silence" is one. The packet process of blocks of packets arriving over consecutive time slots may be described by a MMGBP, [1], as well. More details are presented in a companion paper.

The other important issue in a packet network accommodating packets from sources with different characteristics is that of the allocation of the common facility among the sources. The allocation policy should take into consideration the time constraints imposed on certain packets and the possible monopolization of the common resource by certain sources over long periods; the latter could introduce unacceptable delays to short messages (e.g. consisted of single packets) of interactive communication or control information.

In this paper, we analyze a statistical multiplexing scheme under (dependent) packet arrival processes described by a MMGBP and under a priority policy. The non-i.i.d. MMGBP may be appropriate for the description of complex packet processes while the prioritization may introduce fairness and increased efficiency in the system.

A statistical multiplexer with N packet input processes, each of which is described by a MMGBP has been analyzed in [1], under the first-in first-out (FIFO) service policy. The analysis of the system in [1] (the results of which are presented in the next section) will provide useful results for the analysis of the multiplexing scheme with priorities considered here.

Packets are assumed to arrive through slotted synchronous lines. That is, all packet arrivals are declared at common time instants which coincide with the end of the slots. Discrete time queueing models for statistical multiplexing schemes under non-i.i.d. inputs and without priorities have been analyzed in the past, [1] - [5]. Previous work on statistical multiplexing where packets with different priorities are involved, is heavily based on the assumption of a memoryless packet arrival process (e.g. Poisson), [6]-[8]. Notice that the proposed MMGBP includes simpler processes such as the Bernoulli or the generalized Bernoulli (more than one packet arrivals may occur over the same slot) processes and the first-order Markov process (arrival / no arrival), approximating packet arrivals in bursts or describing the packetized voice traffic. Even under these simple arrival processes and for the priority policy considered in this paper, the corresponding multiplexing scheme has not been analyzed before.

The rest of the paper is organized as follows. In the next section the statistical multiplexer presented in [1] is briefly described and the results from the analysis in [1] are presented. In section III, the multiplexing scheme with priorities is described and the analysis methodology, based on the construction of a system equivalent to the one in [1], is presented. The mean buffer occupancy and the mean packet delay are derived for all packet categories. In section IV, some numerical results on the mean packet delay are presented for the cases considered in section III. Finally, the conclusions of this work appear in the last section.

II. The FIFO Statistical Multiplexer

In this section we describe the statistical multiplexer analyzed in [1] and present the equations derived for the calculations of the mean buffer occupancy and the mean packet delay. This system will be modified to accommodate the priority policies in the next section. By establishing equivalent systems with the one presented briefly in this section, similar equations will be used for the derivation of the queueing results of interest in the next section.

A statistical multiplexer which is fed by N input lines is shown in Fig. 1. The input lines (which are mutually independent) are assumed to be slotted and packet arrivals and service completions are synchronized with the end of the slots. A slot is defined to be the fixed service (transmission) time required by a packet. At most one packet can be served in one slot. The first-in first-out (FIFO) service discipline is adopted. Packets arriving at the same slot are served in a randomly chosen order. The buffer capacity is assumed to be infinite. The packet arrival process associated with line i is defined to be the discrete time process $\{a_j^i\}_{j\geq 0}$, i=1,2,...,N, of the number of packets arriving at the end of the j^{th} slot; $a_j^i=k$, $0\leq k<\infty$, if k packets arrive at the end of the j^{th} slot through input line i.

Figure 1.
The FIFO statistical multiplexer with N inputs.

Let $\{z_j^i\}_{j\geq 0}$, be a finite state Markov chain imbedded at the end of the slots, which describes the state of the input line i. Let $S^i=\{x_0^i, x_1^i, \cdots, x_{M^i-1}^i\}$, $M^i<\infty$, be the state space of $\{z_j^i\}_{j\geq 0}$. It is assumed that the state of the underlying Markov chain determines (probabilistically) the packet arrival process of the corresponding line. That is, if $a^i(x^i): S^i \rightarrow Z_0$, is a probabilistic mapping from S^i into the nonnegative finite integers, Z_0, then the probability that k packets arrive at the buffer at the end of the j^{th} slot is given by $\phi(z_j^i,k)=Pr\{a^i(z_j^i)=k\}$. Furthermore, it is assumed that there is at most one state, x_0^i such that $\phi(x_0^i,0)>0$ and that the rest of the states of the underlying Markov chain result in at least one (but a finite number of) packet arrivals, i.e. $\phi(x_k^i,0)=0$, for $1\leq k\leq M^i-1$. All packet arrivals are assumed to occur at the end of the slots. To avoid instability of the buffer queue it is assumed that there is always one state x_0^i, such as described above.

The expected number of packets in the system is given by, [1],

$$Q = \sum_{\bar{y}\in\bar{S}} W(\bar{y}) \tag{1}$$

where $\bar{S}=S^1xS^2x\cdots xS^N$ and $W(\bar{y})$, $\bar{y}\epsilon\bar{S}$, are the solutions of any $M^1xM^2x\cdots xM^N-1$ linear equations given by

$$W(\bar{y})=\sum_{\bar{x}\in\bar{S}} W(\bar{x})p(\bar{x},\bar{y})+\sum_{\bar{x}\in\bar{S}}(\mu_{\bar{x}}-1)p(\bar{x},\bar{y})\pi(\bar{x}) + \sum_{\bar{x}\in\bar{S}} q_0(\bar{x})p(\bar{x},\bar{y}),\ \bar{y}\epsilon\bar{S} \tag{2a}$$

and the linearly independent equation

$$\sum_{\bar{x}\in\bar{S}}\left[2(\mu_{\bar{x}}-1)W(\bar{x})+2(\mu_{\bar{x}}-1)q_0(\bar{x})+(2+\sigma_{\bar{x}}-3\mu_{\bar{x}})\pi(\bar{x})\right]=0 \tag{2b}$$

where

$$\pi(\bar{x})=\prod_{i=1}^{N}\pi^i(x^i),\quad p(\bar{x},\bar{y})=\prod_{i=1}^{N}p^i(x^i,y^i),\quad q_0(\bar{x})=(1-\lambda)p(\bar{x}_0,\bar{x})$$

$$\mu_{\bar{x}}=\sum_{\nu=1}^{R}\nu g_{\bar{x}}(\nu),\quad \sigma_{\bar{x}}=\sum_{\nu=1}^{R}\nu^2 g_{\bar{x}}(\nu),\quad g_{\bar{x}}(\nu)=Pr\left\{\sum_{i=1}^{N}a^i(x^i)=\nu\right\}$$

and

$$\lambda=\sum_{\bar{x}\in\bar{S}}\mu_{\bar{x}}\pi(\bar{x})<1$$

is the total input traffic which is less than 1 for stability. R is the maximum number of packets which may arrive at the same slot from all N lines; $\pi^i(x^i)$ and $p^i(x^i,y^i)$ are the steady state and the transition probabilities of the Markov chain associated with the i^{th} input line. The mean packet delay is given by using Little's formula, i.e.

$$D=\frac{Q}{\lambda} \tag{3}$$

III. Statistical Multiplexing with priorities

In this section we consider multiplexing schemes under different priority policies. The per slot and line packet arrival processes are described by the MMGBP described in section II.

Case 1:

Consider the statistical multiplexer shown in Fig. 2. The packet arrival process $\{a_j^1\}_{j\geq 0}$ is assumed to be a MMGBP, as described in section II. To avoid monopolization of the facility by long messages (consisted of many packets) which arrive over a single slot, the following service policy is introduced. The first packet of those arrived during a single slot enters a single packet buffer b_1 and it is transmitted in the next slot. The rest of the packets enter an infinite capacity buffer b_2. The server moves to buffer b_2 only if buffer b_1 is empty; it returns to buffer b_1 as soon as this buffer becomes nonempty. This service discipline gives preemptive priority to single packets (over a slot); packets other than the first of a slot are served under a FIFO policy interrupted by new arrivals. This service policy introduces some fairness in the service and it favors single packets.

Clearly, there are two categories of packets, say C_1 and C_2 with different priorities (a smaller subscript indicates higher priority). Packets in C_1 are served (transmitted) first. Thus, the mean delay of packets in C_1, D_1, is equal to 1 (the service time). Service of packets in C_2 is interrupted whenever a packet arrives through line r_1; let D_2 be the mean delay of packets in C_2.

To compute D_2 we consider a FIFO system (shown in Fig. 1) which is equivalent to the one considered here. An equivalent FIFO system is defined as one whose packet arrival processes are identical to those of the system under consideration. If D_{12} denotes the mean packet delay induced by the equivalent FIFO system, then the work conservation law, [8], [9], implies that the following relationship between D_{12}, D_1, and D_2 holds:

$$D_{12}=\frac{\lambda_1 D_1+\lambda_2 D_2}{\lambda_1+\lambda_2} \tag{4}$$

where λ_1 and λ_2 are the per slot packet arrival rates of the packets in C_1 and C_2, respectively. The mean delay of packets in C_2 is given by (4), where λ_1 is equal to $\pi(x\neq x_0)$ (the probability that the line is in any of the packet generating states), $\lambda_2=\lambda_{total}-\lambda_1$ and D_{12} is the mean packet delay of the equivalent FIFO mutiplexer of Fig. 1 computed from equations (1) - (3).

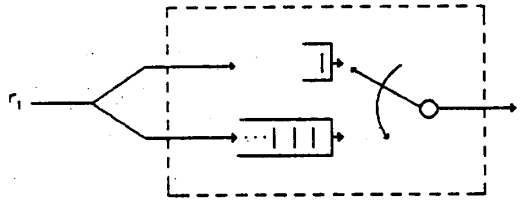

Figure 2.
The statistical multiplexer of Case 1.

Case 2:

Consider the statistical multiplexer shown in Fig. 3. The per input line packet arrival process and the service policy are as in Case 1. The first packet per slot arriving in each of the input lines is given priority by being sent to the infinite buffer b_1; the rest of the packets arriving over the same slot are sent to the infinite buffer b_2. Packets in b_1 have preemptive priority over those in buffer b_2 (as in the previous case). That is, service of the packets in b_2 can start only if buffer b_1 is empty. This service policy avoids monopolization of the facility by either long messages (independently of the generating source) or certain sources (which by nature generate long messages). To compute D_1 and D_2 we proceed as follows.

Calculation of D_1

Consider a FIFO statistical multiplexer (Fig. 1) whose packet arrival process is given by MMGBP's . The underlying Markov chains of these MMGBP's are identical to those associated with the input lines $r_1, \cdots, r_N$ and the total probabilistic mapping

$$a(\bar{x})=\sum_{i=1}^{N} a^i(x^i) \quad , \quad \bar{x}\epsilon\bar{S}$$

is modified to describe the packet arrival process to b_1. That is

$$a_1(\bar{x})=\sum_{i=1}^{N} 1_{(x^i\neq x_0^i)} \quad , \quad \bar{x}\epsilon\bar{S} \tag{5}$$

where x_0^i is the state of line i which generates no packets. Based on (5), the packet generating probabilities $\phi^i(x^i,k)$ are modified to the following

$$\phi^i(x_0^i,0)=1 \quad \text{and} \quad \phi^i(x^i,1)=1 \quad \text{for} \quad x^i\neq x_0^i \tag{6}$$

The mean delay of the packets in D_1 is now computed by applying equations (1) - (3) on the FIFO system with packet arrival processes as determined by (6). The total packet arrival rate λ (used in (3)) is given by

$$\lambda_1=\sum_{i=1}^{N} \pi(x^i\neq x_0^i). \tag{7}$$

Calculation of D_2

The mean delay of packets in buffer b_2 is computed from (4), where λ_1 is given by (7), $\lambda_2=\lambda_{total}-\lambda_1$, and D_{12} is the mean packet delay of the equivalent FIFO multiplexer of Fig. 1, computed from equations (1) - (3).

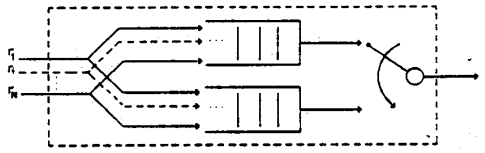

Figure 3.
The statistical multiplexer of Case 2.

IV. Numerical results

In this section some numerical results are derived for each of the priority policies described in the previous section. In the examples considered below it is assumed that the underlying Markov chain associated with any of the input lines has two states, that is $S^i=\{0,1\}$ for the i^{th} line. State 0 is the no-packet generating state (i.e. $a^i(0)=0$); state 1 generates at least one packet, up to a maximum of K^i, with probabilities $\phi^i(1,j)$, $1\leq j\leq K^i$.

As the delay results illustrate, an input traffic process which generates packets clustered around consecutive slots and followed by a period of inactivity, causes significant queueing problems and the induced packet delay is greater than the one induced under better randomized packet arrivals. Since state 1 generates packets and state 0 does not, it makes sense to use the quantity γ^i, where,

$$\gamma^i=p^i(1,1)-p^i(0,1) \tag{8}$$

as a measure of the clusterness of the packet arrival traffic; $p^i(k,j)$ is the probability that the Markov chain associated with line i moves from state k to state j. The value of $\gamma^i=0$ corresponds to a per slot independent packet generation process (generalized Bernoulli process). The clusterness coefficient γ^i and the packet arrival rate λ^i are two important quantities which dramatically affect the delay induced by the multiplexing system. For this reason, each traffic will be characterized by the pair (λ^i,γ^i) and the distribution $\phi^i(1,j)$, $1\leq j\leq K^i$. The rest of the parameters of the MMGBP's associated with each input line are computed from the following equations:

$$\pi^i(1)=\frac{\lambda^i}{\sum_{j=1}^{K^i} j\phi^i(1,j)} \tag{9}$$

$$\pi^i(0) = 1 - \pi^i(1) \quad (10)$$

$$p^i(0,1) = (1-\gamma^i)\pi^i(1) \quad (11)$$

$$p^i(1,1) = \gamma^i + p^i(0,1) \quad (12)$$

$$p^i(1,0) = 1-p^i(1,1) \quad (13)$$

$$p^i(0,0) = 1-p^i(0,1) \quad (14)$$

Case 1:

Consider the multiplexing system of Case 1 with probability distribution

$$\phi^1(1,j) = \begin{cases} .4 & \text{for } j=1 \\ .3 & \text{for } j=2 \\ .2 & \text{for } j=3 \\ .1 & \text{for } j=4 \end{cases}$$

The mean packet delay results D_1, D_2 and D_{12} are shown in Table 1 for various values of $\lambda^1=\lambda$ and $\gamma^1=\gamma$.

Case 2:

Consider the multiplexing system of Case 2 with N = 3 input lines, probability distributions as in Case 1 and parameters $\lambda^1=\lambda^2=\lambda^3=\lambda/3$ and $\gamma^1=\gamma^2=\gamma^3=\gamma$. The mean packet delay results D_1, D_2 and D_{12} are shown in Table 2 for various values of λ and γ.

V. Conclusions

In this paper a statistical multiplexing scheme under a priority policy has been analyzed. The per input line packet arrival processes are described by the Markov Modulated Generalized Bernoulli Process (MMGBP) defined in section II. The investigated packet multiplexing scheme has, in essense, served as an illustration of a general methodology which utilizes the results of the analysis of a first-in first-out (FIFO) equivalent multiplexing scheme under identical packet arrival processes. For the cases considered in this paper, the MMGBP describing the per line packet arrival process is transformed into another process of the same type (i.e. a MMGBP), which incorporates the priority policy. Thus, equivalent FIFO multiplexing systems are constructed with inputs described by a MMGBP, as well. The previous property of the MMGBP facilitates the analysis of the multiplexing systems under priorities through the formulation of equivalent FIFO systems. On the other hand, the MMGBP can serve as a model for a wide class of complex packet arrival processes and thus, facilitate the appropriate description and the analysis of many practical systems.

λ	γ	D_1	D_{12}	D_2
.90	.5	1.000	18.500	36.000
.90	.3	1.000	12.786	24.571
.90	.0	1.000	8.500	16.000
.70	.5	1.000	6.833	12.667
.70	.3	1.000	4.928	8.857
.70	.0	1.000	3.500	6.000

Table 1
Mean packet delay results for Case 1.

λ	γ	D_1	D_{12}	D_2
.90	.5	1.818	27.500	53.181
.90	.3	1.506	18.357	35.208
.90	.0	1.273	11.500	21.727
.70	.5	1.538	9.167	16.795
.70	.3	1.333	6.373	11.413
.70	.0	1.179	4.278	7.376

Table 2
Mean packet delay results for Case 2.

References

1. I. Stavrakakis, "A Statistical Multiplexer for Packet Networks", IEEE Transactions on Communications (submitted).
2. H. Heffes, D. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Analysis", IEEE Journal on Selected Areas in Communications, Vol. SAC-4, No. 6, Sept. 1986.
3. D. Towsley, "The Analysis of a Statistical Multiplexer with Nonindependent Arrivals and Errors", IEEE Transactions on Communications, Vol. COM-28, No. 1, Jan. 1980.
4. H. Bruneel, "Queueing Behavior of Statistical Multiplexers with Correlated Inputs", IEEE Transactions on Communications, Vol. COM-36, No. 12, Dec. 1988.
5. A. Viterbi, "Approximate Analysis of Time Synchronous Packet Networks", IEEE Journal on Selected Areas in Communications, Vol. SAC-4, No. 6, Sept. 1986.
6. Chung-Yin Lo, "Performance Analysis and Application of a Two Priority Packet Queue", AT&T Technical Journal, Vol.66, Issue 3, May/June 1987.
7. M. Hluchyj, C. Tsao R. Boorstyn, Performance Analysis of a Preemptive Priority Queue with Applications to Packet Communication Systems", The Bell System Technical Journal, Vol.62, No.10, Dec. 1983.
8. G. Barberis, "A Useful Tool in the Theory of Priority Queueing", IEEE Transactions on Communications, Vol. COM-28, No. 9, Sep. 1980.
9. D. Heyman, M. Sobel, "Stochastic Models in Operations Research, Vol.1", McGraw-Hill, 1982.

IOANNIS STAVRAKAKIS

He received his Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1983 and his Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in 1988.

In 1988 he joined the Department of Electrical Engineering and Computer Science, University of Vermont, Burlington, where he is presently an Assistant Professor. His research interests include multi-user communication theory, stochastic processes, queueing theory and system performance evaluation.