Bounds on Buffer Overflow Probabilities in Communication Systems

Kurt Geihs *
Department of Computer Science, JWG-Universität, D-6000 Frankfurt, Germany

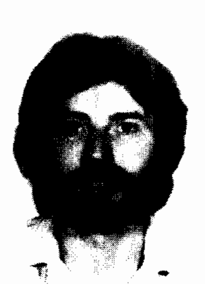
H. Kobayashi **
The Science Institute, IBM Japan Ltd., 5-1, Azabudai 3-chome Minato-ku, Tokyo 106, Japan

Received 20 October 1980
Revised 1 April 1982

A model to analyze the buffer behaviour in a multiplexor is derived, based on the analogy between the buffer occupancy in a discrete time model of multiplexing and the waiting time of a GI/G/1 queueing system.

The bounding techniques developed earlier by Kingman and Ross are extended to the discrete time model. Simple and useful bounds are obtained for the buffer overflow probabilities under general assumptions concerning incoming message traffic characteristics. Numerical examples are presented and compared with other methods.

Keywords: Buffer Overflow, Concentrator, GI/G/1 Queue, Multiplexor, Overflow Probability, Kingman–Ross Bounds, Sequential Analysis, Wald's Identity, IFR, DFR.



Kurt Geihs was born in Germany in 1955. He received his Dipl.-Inform. degree in Computer Science from the Technische Hochschule Darmstadt, Germany, in 1980. From 1980 to 1981 he spent one year at the University of California, Los Angeles, where he received a M.S. in Computer Science. Presently, he is working toward his Ph.D. degree at the University of Frankfurt, Germany. His research interests include computer communications, modeling and analysis of computer systems and distributed systems.

* While the author was with the Department of Computer Science, the Technical University of Darmstadt, West Germany.

** The work was supported by the Alexander von Humboldt Foundation, West Germany, while the author was with the Technical University of Darmstadt.

North-Holland Publishing Company
Performance Evaluation 2 (1982) 149–160

1. Introduction

The streams of messages generated from terminals or data from computers are usually not steady flows: they are quite often sporadic or 'bursty'. In order to enhance the utilization of communication links, such techniques as statistical multiplexing, concentration and packet-switching multiplexing are commonly adopted in a majority of teleprocessing systems and computer communication networks [2].

In such devices several independent arrival streams with large variation in time are combined to form a single outgoing flow of a more regular nature. This transformation can be accomplished by interposing buffer storage between the set of incoming links and the output link (Fig. 1). The buffer space should be sufficiently large to accommodate a large backlog of messages or data units which may occasionally develop. The buffer's capacity and its allocation strategy are important

Hisashi Kobayashi is Director of the Science Institute, a newly established research organization of IBM Japan Corporation in Tokyo. Formerly he was Senior Manager of VLSI Design at the IBM T.J. Watson Research Center, Yorktown Heights, New York. He authored *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology* (Addison-Wesley, Reading, MA, 1978) and over 60 papers in the fields of system performance analysis, queueing theory, data transmission theory and signal processing. Currently he is Chairman of IFIP Working Group 7.3 on "Computer System Modeling", Associate Editor of the *IEEE Trans. on Information Theory*, and Editor-in-Chief of the *Performance Evaluation* Journal.

Among the various honors he has received are the following: IEEE Fellow (1977), the Humboldt Award (1979) and the Silver Core from IFIP (1980). He is listed in Who's Who in America since 1982.

He received B.S. and M.S. in Electrical Engineering from the University of Tokyo in 1961 and 1963, respectively, and worked for Toshiba Electric Company from 1963 till 1965. He received his M.A. and Ph.D. degrees from Princeton University in 1966 and 1967, respectively, and has been with the IBM Corporation since 1967. He has taught at UCLA, University of Hawaii, Stanford University, Technische Hochschule Darmstadt in Germany, and at Université Libre de Bruxelles in Belgium.

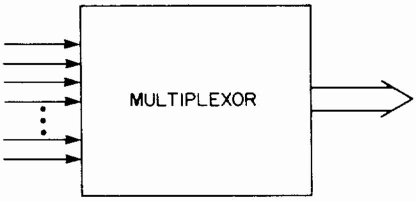


Fig. 1.

to the cost-performance tradeoffs in designing such a communication system. Therefore, the analysis of buffer behavior, particularly of the buffer overflow probability, is of considerable importance. There are a number of papers reported (see, for instance, [2, Chapter 8] and the references therein). But the traffic assumptions are overly restrictive – arrivals are Poisson, or the output link capacity is constant – and the solution usually involves a large number of system equations to be solved numerically.

In this paper we present very simple upper and lower bounds of the buffer occupancy distribution under fairly general assumptions concerning the characteristics of incoming traffic and the output rate of the multiplexor. An upper bound on the buffer overflow probability is also derived. These probabilities all take a simple form z_0^n, where z_0 is the root of the characteristic equation determined by the input and output statistics.

Our model formulation follows that of Wyner [3]. We derive very simple and useful upper and lower bounds based on the results of Kingman [4] and Ross [5]. In Section 2 we briefly review the Kingman–Ross bounds and their discrete time analogues. In Section 3 we derive the fundamental inequalities for buffer analysis by using Wald's identity of sequential decision. Applications of the Kingman–Ross results to the buffer problem were earlier reported by Kobayashi and Konheim [6] and Kobayashi [1], but the rigorous derivation of the discrete time analogue of the Kingman–Ross bounds is given here for the first time. In Section 4 we discuss some numerical examples and compare them with earlier work by Wyner [3] and Chu [7].

2. Statement of the problem

2.1. The Kingman–Ross inequality

The mathematical formulae describing the behavior of a buffer are analogous to equations for the waiting time in a GI/G/1 queueing system, as will be shown in Section 2.2. Therefore, we briefly review some of the GI/G/1 results pertinent to the present paper.

Lindley [8] showed that the waiting time sequence $\{w_k\}$ satisfies the following recurrence equation:

$$w_k = \max\{0, w_{k-1} + u_k\} \tag{2.1}$$

where

w_k = waiting time of the kth customer ($w_0 = 0$),

$u_k = x_{k-1} - t_k$,

x_k = service time of the kth customer

and

t_k = interarrival time between $(k-1)$st and the kth customer.

Starting with relation (2.1), Lindley found the stationary waiting time distribution

$$F_w(x) = \lim_{k\to\infty} \mathrm{P}\{w_k \leqslant x\} \tag{2.2}$$

as the solution of a Wiener–Hopf type integral equation. Solving this equation generally requires techniques from the theory of complex variables and contour integration. Therefore, we cannot obtain an analytic expression in terms of known quantities for the waiting time distribution or even for the mean waiting time in a GI/G/1 queueing system. For this reason, there is an increasing interest and need for developing approximations and bounding techniques [4, 5, 9, 10]. The following results obtained by Kingman [4] and Ross [5] have a fundamental bearing upon the present paper. Kingman proved, by using Kolmogorov's inequality for martingales, that the tail of the waiting time distribution

$$\bar{F}_w(t) = \lim_{k\to\infty} \mathrm{P}\{w_k > t\}$$

is bounded by

$$a\,\mathrm{e}^{-\theta_0 t} \leqslant \bar{F}_w(t) \leqslant \mathrm{e}^{-\theta_0 t}. \tag{2.3}$$

Ross improved the Kingman's upper bound by applying Wald's fundamental identity of sequential analysis, and obtained

$$a\,\mathrm{e}^{-\theta_0 t} \leqslant \bar{F}_w(t) \leqslant b\,\mathrm{e}^{-\theta_0 t} \tag{2.4}$$

where the parameters a, b and θ_0 in (2.3) and (2.4) are given by

$$a = \inf_{s\geqslant 0} D(s), \tag{2.5}$$

$$b = \sup_{s \geq 0} D(s), \tag{2.6}$$

$$D(s) = 1/\mathrm{E}\left[e^{\theta_0(u-s)} \mid u > s\right], \tag{2.7}$$

$$\theta_0 \text{ is the characteristic root of } \int_{-\infty}^{\infty} e^{\theta y} \, dH(y) = 1 \text{ such that } \theta > 0 \tag{2.8}$$

where

$$H(s) = \lim_{k \to \infty} \mathrm{P}\{u_k \leq s\}. \tag{2.9}$$

2.2. The buffer behavior model

In this paper we assume that the time axis is divided into a sequence of 'slots' or 'intervals'. This is certainly a proper model to assume when the communication system is *synchronous*, i.e., all events are allowed to occur only at definite, regularly spaced time points. Examples are synchronous time division multiplexing and the slotted ALOHA multiplexing scheme [2,10]. Even for *asynchronous* systems we often find it computationally convenient to deal with discrete time systems, when we must eventually numerically evaluate the performance. The size of the slot certainly affects the accuracy of the model, but here we simply assume that this time unit is properly chosen.

Let a_k be the total number of messages or data units arriving to the multiplexor in the kth time slot, and b_k be the maximum number of data units that can leave the multiplexor in the kth time slot. Here $\{a_k\}$ and $\{b_k\}$ are random variables. Ordinarily b_k is a constant parameter determined by the link capacity of the multiplexor output. But in our model it is not necessary to assume that b_k is constant. A possible application of such a model will be the multiplexing operation performed in a satellite ground station that accesses the satellite transponder based on the demand assigned TDMA (time division multiple access): The satellite channel is dynamically partitioned among 'traffic bursts of variable durations transmitted from different earth stations. Hence the output bandwidth allocated to the multiplexor of a given earth station varies in time. Another example of the variable b_k will be found when the multiplexor combines different types of traffics, say voices and data. Depending on the priority assignment, the net output rate of the multiplexor of data channels may vary with the voice traffic level.

Let us assume that the sequence of differences $\{c_k\}$ defined by

$$c_k = a_k - b_k, \quad k = 1, 2, 3, \ldots \tag{2.10}$$

are independent, identically distributed (i.i.d.) random variables. Let y_k denote the buffer occupancy at the end of the kth time slot, and let L be the capacity of the buffer. Then the sequence $\{y_k\}$ satisfies the following relation [3,6]:

$$y_k = \begin{cases} 0, & y_{k-1} + c_k \leq 0, \\ y_{k-1} + c_k, & 0 \leq y_{k-1} + c_k \leq L, \\ L, & L \leq y_{k-1} + c_k. \end{cases} \tag{2.11}$$

Now consider the limit case where the buffer capacity is infinitely large. If we denote the corresponding buffer occupancy by $\{x_k\}$, we obtain

$$x_k = \max\{0, x_{k-1} + c_k\}. \tag{2.12}$$

Note that (2.1) and (2.12) take exactly the same mathematical form except that the variables $\{x_k\}$ take on only integers, whereas $\{w_k\}$ is a sequence of real variables. Kobayashi and Konheim [6] applied the Kingman–Ross inequality to derive lower and upper bounds on

$$\bar{F}_x(n) = \lim_{k \to \infty} \mathrm{P}\{x_k > n\}, \quad n = 0, 1, 2, \ldots. \tag{2.13}$$

The discrete analogue of (2.4) should be [1]

$$A z_0^{-n} \leq \bar{F}_x(n) \leq B z_0^{-n} \tag{2.14}$$

where

$$A = \inf_{m \geq 0} D(m), \tag{2.15}$$

$$B = \sup_{m \geq 0} D(m), \tag{2.16}$$

$$D(m) = 1/\mathrm{E}\left[z_0^{c_k - m} \mid c_k > m\right], \tag{2.17}$$

$$z_0 \text{ is the characteristic root of } H(z) = 1 \text{ such that } z_0 > 1 \tag{2.18}$$

and

$$H(z) = \mathrm{E}[z^{c_k}]. \tag{2.19}$$

Using the upper bound of (2.14) and the relation

$$\lim_{k \to \infty} \mathrm{P}\{y_k = L\} \leq \sum_{n=L}^{\infty} \lim_{k \to \infty} \mathrm{P}\{x_k = n\} \tag{2.20}$$

[1] There is an error in equations (129) and (130) of [6]: the condition $\alpha \geq n$ should be $\alpha > n$.

we obtain

$$\mathrm{P}\{\text{The buffer of capacity } L \text{ is full}\} \leqslant B z_0^{-(L-1)}. \quad (2.21)$$

As mentioned earlier, these results were not explicitly proven in [6] and therefore require formal derivations. By doing so we have in fact discovered some minor erros in the formulae presented in [6].

3. Derivations of the bounds

3.1. Model assumptions and preliminaries

In our model the variables $\{c_k\}$ where $c_k = a_k - b_k$ are assumed to be i.i.d. random variables. In order for a system to be stable the following condition must hold:

$$\mathrm{E}[c_k] < 0. \quad (3.1)$$

We also assume

$$\mathrm{P}\{c_k > 0\} > 0, \quad (3.2)$$

which means that sometimes the amount of arriving data, a_k, exceeds the system capacity, b_k. Otherwise there would be no need for buffering. We assume that the buffer is initially empty, i.e.,

$$x_0 = 0. \quad (3.3)$$

The buffer occupancy at the end of the first time slot is (assuming an infinite buffer capacity)

$$x_1 = \max\{0, x_0 + c_1\} = \max\{0, c_1\}. \quad (3.4)$$

By repeating the recurrence relation (2.12) we write x_k in terms of $c_1, c_2, \dots, c_k$ [8]:

$$x_k = \max\{0, c_k, c_k + c_{k-1}, \dots, c_k + c_{k-1} + \cdots + c_2 + c_1\}. \quad (3.5)$$

Since all c_i's are i.i.d., we replace c_i by c_{k+1-i} and obtain the following expression, which is stochastically equivalent to (3.5):

$$x_k = \max\{0, c_1, c_1 + c_2, c_1 + c_2 + c_3, \dots, c_1 + c_2 + \cdots + c_k\}. \quad (3.6)$$

Thus the complementary distribution function of x_k is given by

$$\bar{F}_{x_k}(n) = \mathrm{P}\{x_k > n\} = \mathrm{P}\left\{\sum_{i=1}^{j} c_i > n \text{ for some } j = 1, 2, \dots, k\right\}. \quad (3.7)$$

In the equilibrium state, i.e., in the limit $k \to \infty$ the last expression becomes

$$\bar{F}_x(n) = \mathrm{P}\left\{\sum_{i=1}^{j} c_i > n \text{ for some } j = 1, 2, \dots\right\}. \quad (3.8)$$

Recall that the random variables $\{c_i\}$ are integer-valued. The right-hand side of expression (3.8) suggests us to investigate a discrete-valued version of the theory of sequential analysis [11].

3.2. Wald's fundamental identity

Let us first examine roots of the equation

$$H(z) = 1 \quad (3.9)$$

where $H(z)$ was defined by (2.19). From the definition of the generating function we readily see that $z_1 = 1$ is one of the roots. In Appendix A we show that under the assumptions (3.1) and (3.2) there is exactly one root $z_0 > 1$ that satisfies $H(z_0) = 1$. The next step is to apply the following identity, which is an integer-valued version of Wald's fundamental identity:

$$\mathrm{E}\left[z^{\sum_{i=1}^{j} c_i} H(z)^{-j}\right] = 1. \quad (3.10)$$

for all z such that $H(z) \geqslant 1$. Here the integer j is the variable that signifies the time that the running sum

$$S_i = c_1 + c_2 + \cdots + c_i \quad (3.11)$$

falls outside of some prefixed interval, say $[\underline{S}, \bar{S}]$, for the first time, i.e.,

$$j = \min\{i: S_i \notin [\underline{S}, \bar{S}]\}. \quad (3.12)$$

In the theory of sequential analysis the variable j is called the 'stopping time', because the sample size is not fixed, but in step i a new data c_i is taken and examined to test whether a given hypothesis should be accepted or rejected, or whether the test is to continue. In Appendix B we derive the equation (3.10). Note that Wald's identity was originally derived for continuous variables [11,12].

By setting $z = z_0$ in (3.10) we obtain

$$\mathrm{E}\left[z_0^{S_j}\right] = 1. \quad (3.13)$$

In our problem we choose the prefixed interval as

$$[\underline{S}, \bar{S}] = (-\infty, n] \quad (3.14)$$

where n is a fixed positive integer. Thus, the stopping time j is a variable that is determined by the threshold value n and the observed data $\{c_i\}$. Noting that at time j the running sum S_j exceeds n for the first time, we can rewrite (3.13) as

$$1 = \mathrm{E}\left[z_0^{S_j} | S_j > n\right] \mathrm{P}\{S_j > n\}$$
$$= z_0^n \mathrm{E}\left[z_0^{S_j - n} | S_j > n\right] \bar{F}_x(n) \quad (3.15)$$

where the last term was obtained by using (3.8).

In order for $S_j > n$ to hold, there must be *some* integer $m \geqslant 0$ such that

$$S_j - n = c_j - m. \quad (3.16)$$

Since the conditional distribution of $S_j - n$ given that $j = k$, $S_j = n - m$ and $S_j > n$ is just the conditional distribution of $c_k - m$ given that $c_k > m$, it follows by conditioning on j and $S_j - 1$ that

$$\inf_{m \geqslant 0} \mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right] \leqslant \mathrm{E}\left[z_0^{S_j - n} | S_j > n\right] \leqslant \sup_{m \geqslant 0} \mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right]. \quad (3.17)$$

Hence from (3.15) and (3.17) we find

$$\bar{F}_x(n) z_0^n \inf_{m \geqslant 0} \mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right] \leqslant 1 \leqslant \bar{F}_x(n) z_0^n \sup_{m \geqslant 0} \mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right], \quad (3.18)$$

from which we finally derived inequality (2.14), i.e.,

$$A z_0^{-n} \leqslant \bar{F}_x(n) \leqslant B z_0^{-n} \quad (3.19)$$

where

$$A = \inf_{m \geqslant 0} 1 / \mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right] \quad (3.20)$$

and

$$B = \sup_{m \geqslant 0} 1 / \mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right]. \quad (3.21)$$

By following the argument similar to Ross [5], we can derive a somewhat weaker but probably more computable inequality. For this purpose we write

$$\mathrm{E}\left[z_0^{c_1 - m} | c_1 > m\right] = \mathrm{E}\left[z_0^{a_1 - (b_1 + m)} | a_1 > b_1 + m\right] = \mathrm{E}\left[z_0^{a_1 - m'} | a_1 > m'\right]. \quad (3.22)$$

Thus, we find

$$A^* z_0^{-n} \leqslant A z_0^{-n} \leqslant \bar{F}_x(n) \leqslant B z_0^{-n} \leqslant B^* z_0^{-n} \quad (3.23)$$

where

$$A^* = \inf_{m > 0} 1 / \mathrm{E}\left[z_0^{a_1 - m} | a_1 > m\right] \quad (3.24)$$

and

$$B^* = \sup_{m > 0} 1 / \mathrm{E}\left[z_0^{a_1 - m} | a_1 > m\right]. \quad (3.25)$$

3.3. Finite buffer capacity

So far we have assumed the infinite buffer case. For the finite buffer case we denote the buffer occupancy sequence by $\{y_k\}$, as defined in (2.11), and compare it with $\{x_k\}$ of (2.12). From the recurrence (2.10) we have

$$y_1 = \min\{L, \max\{0, c_1\}\} \leqslant \max\{0, c_1\} = x_1. \quad (3.26)$$

If we assume that

$$y_k \leqslant x_k \quad \text{for some } k, \quad (3.27)$$

then for $k + 1$

$$y_{k+1} = \min\{L, \max\{0, y_k + c_{k+1}\}\} \leqslant \min\{L, \max\{0, x_k + c_{k+1}\}\} \leqslant \max\{0, x_k + c_{k+1}\} = x_{k+1}. \quad (3.28)$$

Thus we have shown by mathematical induction that

$$y_k \leqslant x_k \quad \text{for all } k. \quad (3.29)$$

Then the distribution of the variables $\{y_k\}$ and $\{x_k\}$ are related as follows:

$$\mathrm{P}\{y_k \geqslant 0\} \leqslant \mathrm{P}\{x_k \geqslant n\} \quad \text{for all } k. \quad (3.30)$$

Since the maximum value that y_k can take is L, it must follow that

$$\mathrm{P}\{y_k = L\} \leqslant \mathrm{P}\{x_k \geqslant L\} \quad \text{for all } k. \quad (3.31)$$

Let us denote the distributions of the random variables y_k and x_k in the limit $k \to \infty$ by $p_L(n)$ and $p_\infty(n)$, respectively. Then the last equation becomes

$$p_L(L) \leqslant \sum_{n=L}^{\infty} p_\infty(n) = \bar{F}_x(L-1). \quad (3.32)$$

Hence, using (3.23) and (3.32) we obtain an upper bound for the probability that the buffer of capacity L becomes full:

$$p_L(L) \leqslant B z_0^{-(L-1)} \leqslant B^* z_0^{-(L-1)}. \quad (3.33)$$

Now we define the quantity $P_{\text{overflow}}(L)$ as the percentage of lost data due to overflow [3,7]:

$$P_{\text{overflow}}(L) = \frac{\text{offered load} - \text{carried load}}{\text{offered load}} = Q(L) / \mathrm{E}[a_k] \quad (3.34)$$

where $Q(L)$ is the expected amount of data lost per time slot, i.e.,

$$Q(L) = \sum_{n=0}^{L} p_L(n)\left(\sum_{k=0}^{\infty} kh_{L-n+k}\right). \tag{3.35}$$

Here h_j is the coefficient of the term z^j in the polynominal $H(z)$ of (2.7), that is to say

$$h_j = \mathrm{P}\{c_k = j\}. \tag{3.36}$$

Using the upper limits

$$p_L(n) \leqslant \sum_{i=n}^{\infty} p_\infty(i) = \bar{F}_x(n-1) \quad \text{for } n \geqslant 1 \tag{3.37}$$

and

$$p_L(0) \leqslant 1, \tag{3.38}$$

we obtain an upper bound on $Q(L)$:

$$Q(L) \leqslant Q_+(L)$$
$$\triangleq \sum_{k=0}^{\infty} kh_{L+k} + \sum_{n=1}^{L}\left(Bz_0^{-(n-1)} \sum_{k=0}^{\infty} kh_{L-n+k}\right). \tag{3.39}$$

Thus from (3.24) we find that

$$P_{\text{overflow}}(L) \leqslant Q_+(L)/\mathrm{E}[a_k]. \tag{3.40}$$

3.4. The function D(n)

In order to evaluate the constants A and B of (3.20) and (3.21) we must find the infimum and supremum of the following function:

$$D(n) = \frac{1}{\mathrm{E}\left[z_0^{c_k - n} \mid c_k > n\right]}. \tag{3.41}$$

We rewrite this equation as

$$D(n) = \frac{1}{z_0 \mathrm{E}\left[z_0^{c_k - n - 1} \mid c_k \geqslant n+1\right]}, \tag{3.42}$$

from which we readily find an upper limit on $D(n)$:

$$D(n) \leqslant \frac{1}{z_0} < 1. \tag{3.43}$$

A similar argument can be applied to bounding B^* of (3.25); indeed B^* is also bounded by z_0^{-1}, i.e.,

$$B \leqslant B^* \leqslant z_0^{-1} < 1, \tag{3.44}$$

which in turn implies

$$\bar{F}_x(n) \leqslant z_0^{-(n+1)}. \tag{3.45}$$

Hence, from (3.33) we find

$$p_L(L) \leqslant z_0^{-L}. \tag{3.46}$$

In order to establish a lower bound on A (or A^*) it is necessary to make some assumptions concerning the distributional form of the random variables $\{c_k\}$ (or $\{a_k\}$). Therefore, we make the following definition of an IFR (increasing failure rate) and DFR (decreasing failure rate) discrete distribution function. [2]

Definition 1. The distribution of a discrete random variable z is called *IFR* (*DFR*) if and only if

$$\bar{F}(m+n)/\bar{F}(n)$$

is nonincreasing (nondecreasing) in n for all $m \geqslant 0$, where $\bar{F}(n) = 1 - F(n) = \mathrm{P}\{z > n\}$.

Now, for any function $u(z)$ such that $\mathrm{E}[u(z)] < \infty$, and for the discrete distribution $F(m)$ (with the underlying variable z), we find the formula

$$\mathrm{E}[u(z)] = \sum_{m=-\infty}^{\infty} (u(m+1) - u(m))\bar{F}(m). \tag{3.47}$$

Then it is not difficult to show that

$$\mathrm{E}\left[z_0^{c_k - n} \mid c_k > n\right] =$$
$$= z_0 + \sum_{m=1}^{\infty}\left(z_0^{m+1} - z_0^{m}\right)\frac{\bar{F}_c(m+n)}{\bar{F}_c(n)}. \tag{3.48}$$

Since $z_0 > 1$, the above expression is a nonincreasing (nondecreasing) function of n if $F_c(n)$ is IFR (DFR). Hence, if $F_c(n)$, the distribution of the random variables $\{c_k\}$, is IFR, then we find that

$$A = \inf_{n \geqslant 0} D(n) = D(0) = 1/\mathrm{E}\left[z_0^{c_k} \mid c_k > 0\right] \tag{3.49}$$

and

$$A = \sup_{n \geqslant 0} D(n) = \lim_{n \to N} D(n) \tag{3.50}$$

where N is the maximum possible integer that the random variables $\{c_k\}$ can take on, i.e.,

$$N = \sup\{n\colon \bar{F}_c(n) > 0\}. \tag{3.51}$$

The same arguments apply to the DFR case just be replacing ‘decreasing’ by ‘increasing’. Then we

[2] Note that our definition is somewhat different from the definition in [13]. We assume that z is an integer-valued variable, but not necessarily non-negative.

obtain for $F_c(n)$ being DFR:

$$A = \inf_{n \geq 0} D(n) = \lim_{n \to N} D(n) \tag{3.52}$$

and

$$B = \sup_{n \geq 0} D(n) = D(0) = 1/\mathrm{E}\left[z_0^{c_k} \mid c_k > 0\right]. \tag{3.53}$$

It goes without saying that the same techniques can be applied to the computation of A^* and B^*, when the arrival distribution is either IFR or DFR.

4. Examples

4.1. Poisson arrivals and constant rate outputs

Suppose that the arrival process is Poisson with rate λ (data units/time slot), and that the concentrator sends out μ (data units/time slot).[3] Then the distributions of the random variables $\{a_k\}$ and $\{b_k\}$ are given by the following $\{f_n\}$ and $\{g_n\}$, respectively:

$$f_n = \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} \quad \text{for all } n \geq 0 \tag{4.1}$$

and

$$g_n = \begin{cases} 1 & \text{for } n = \mu, \\ 0 & \text{for } n \neq \mu. \end{cases} \tag{4.2}$$

The corresponding generating functions are obtained as

$$F(z) = \mathrm{E}[z^{a_k}] = \mathrm{e}^{\lambda(z-1)} \tag{4.3}$$

and

$$G(z) = \mathrm{E}[z^{b_k}] = z^{\mu}. \tag{4.4}$$

Thus,

$$H(z) = F(z)G(z^{-1}) = \mathrm{e}^{\lambda(z-1)} z^{-\mu}. \tag{4.5}$$

From (4.5) and assumption (3.1) we find that $H(z)$ has a negative derivative at $z = 1$, i.e.,

$$H'(1) = \lambda - \mu = \mathrm{E}[c_k] < 0. \tag{4.6}$$

To calculate z_0 we set $H(z_0)$ to unity, and we define the traffic intensity $\rho = \lambda/\mu$. Then we find

$$\mathrm{e}^{\rho(z_0 - 1)} = z_0. \tag{4.7}$$

Thus, the root z_0 depends on λ and μ only through the traffic intensity ρ. As no analytic solution[4] is found for (4.7), a numerical root finding algorithm, such as Newton's iteration algorithm, should be applied.

It is not difficult, although somewhat tedious, to show that the Poisson distribution $\{f_n\}$ of (4.1) is indeed IFR. Therefore, the distribution $\{h_n\}$, which is a shifted version of $\{f_n\}$ (i.e., $h_n = f_{n+\mu}$, $n \geq -\mu$), is also shown to be IFR. Hence, the function $D(N)$ is monotone nondecreasing in n, and we find

$$A = D(0) = z_0^{\mu} \frac{\mathrm{e}^{\lambda} - \sum_{i=0}^{\mu} \frac{\lambda^i}{i!}}{\mathrm{e}^{\lambda z_0} - \sum_{i=0}^{\mu} \frac{(z_0 \lambda)^i}{i!}} \tag{4.8}$$

and

$$B = \lim_{n \to \infty} D(n) = 1/z_0, \tag{4.9}$$

where the expression (4.8) is obtained by using (3.48). Similarly we obtain

$$A^* = \frac{\mathrm{e}^{\lambda} - 1}{\mathrm{e}^{\lambda z_0} - 1} \tag{4.10}$$

and

$$B^* = 1/z_0. \tag{4.11}$$

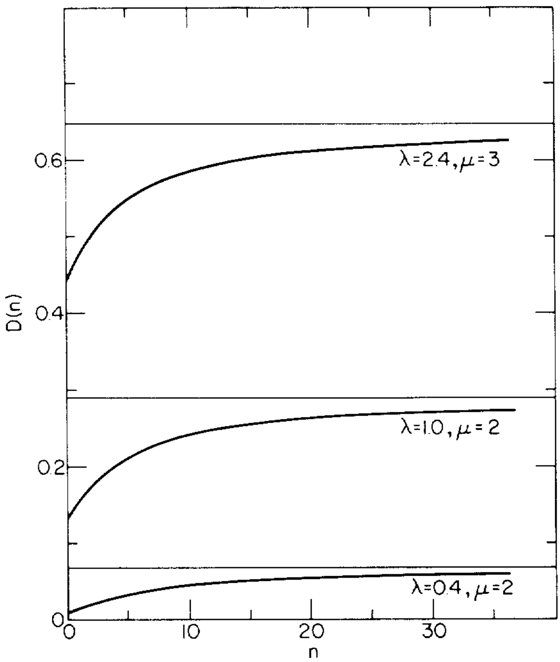


Fig. 2. Poisson arrivals and constant output. The function $D(n)$ (bounded by $1/z_0$) vs. n for various values of λ and μ.

[3] μ is a positive integer, but λ does not need to be an integer.

[4] [6, Eq. (144)] is incorrect: it gives a root less than unity.

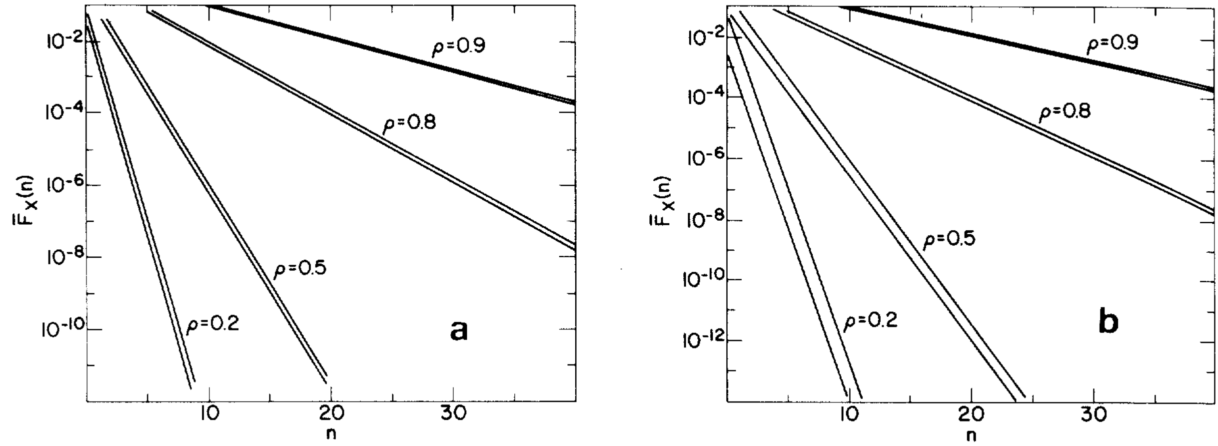


Fig. 3. Poisson arrivals and constant rate output. Upper and lower bounds on $\mathbf{P}\{x_k > n\}$ vs. n for various values of the traffic intensity: $\rho = 0.2$, 0.5, 0.8 and 0.9. (a) $\mu = 1$; (b) $\mu = 3$.

Fig. 2 shows the curves of $D(n)$ and the corresponding upper limits $1/z_0$ for three different examples. The lower and upper bounds on $\bar{F}_x(n)$ are plotted in fig. 3(a) and (b) for various values of traffic intensities when the output trunk capacity $\mu = 1$ and $\mu = 3$, respectively. Note that the distribution $\bar{F}_x(n)$ depends primarily on $\rho = \lambda/\mu$, as does z_0. More details concerning the computation of z_0 and $D(n)$ can be found in [14].

4.2. Geometric arrivals and constant rate outputs

Let the number of arrivals per time slot be geometrically distributed with parameter $r < 1$, i.e.,

$$f_n = \mathrm{P}\{a_k = n\} = (1-r)r^n. \tag{4.12}$$

The corresponding generating function is given by

$$F(z) = \frac{1-r}{1-rz}, \quad |z| < 1/r. \tag{4.13}$$

We assume, as in Section 4.1, that the concentrator sends out data with rate μ (data units/time slot). Then we obtain

$$H(z) = \frac{(1-r)}{(1-rz)} z^{-\mu}, \quad 0 < |z| < 1/r, \tag{4.14}$$

from which the characteristic equation for z_0 can be written as

$$z_0^{\mu+1} - \frac{1}{r} z_0^{\mu} + \frac{1-r}{r} = 0. \tag{4.15}$$

It can be easily shown that there exists a unique root z_0 such that

$$1 < z_0 < 1/r. \tag{4.16}$$

By referring to (3.24) and (3.25) we evaluate

$$\mathrm{E}\left[z_0^{a_k - n} | a_k > n\right] = \frac{(1-r)z_0}{1-rz_0}, \tag{4.17}$$

which is independent of n. This surprisingly simple result is due to the memoryless property of the geometric distribution. Note also that this distribution is both IFR and DFR. Therefore, the supremum and infimum of (4.17) are equal, and we find $A^* = B^*$, which in turn implies that the function $D(n)$ is also independent of n, i.e.,

$$B^* = B = A = A^* = \frac{1-rz_0}{(1-r)z_0}. \tag{4.18}$$

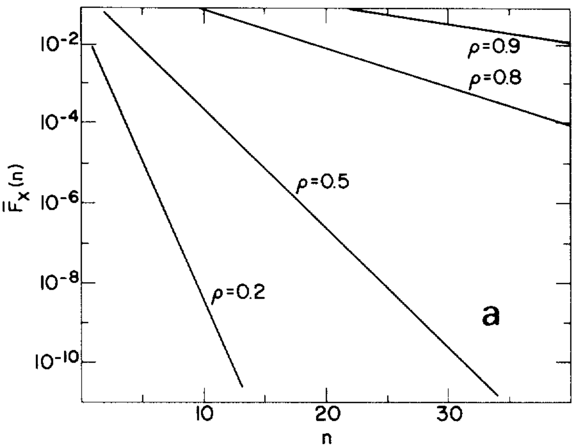
We therefore have the following exact expression for $\bar{F}_x(n)$:

$$\bar{F}_x(n) = \lim_{k\to\infty} \mathrm{P}\{x_k > n\} = \frac{1-rz_0}{1-r} z_0^{-(n+1)}. \tag{4.19}$$

Note that the simple form of (4.19) holds, even when the output sequence $\{b_k\}$ has a general distribution. That is, the multiplexing system with geometric arrivals and variable output has the buffer occupancy distribution also given by (4.19). Of course the distribution of the b_k's influences the characteristic root z_0. It will be also worthwhile mentioning that this constant value of $D(n)$ is strictly smaller than $1/z_0$.

Fig. 4(a) and (b) shows $\bar{F}_x(n)$ for transmission rates $\mu = 1$ and $\mu = 3$. The z_0's were derived as follows. For $\mu = 1$, we readily obtain $z_0 = (1-r)/r$. In case of $\mu = 3$ the characteristic equation becomes

$$z_0^4 + \frac{1}{r} z_0^3 + \frac{1-r}{r} = 0. \tag{4.20}$$



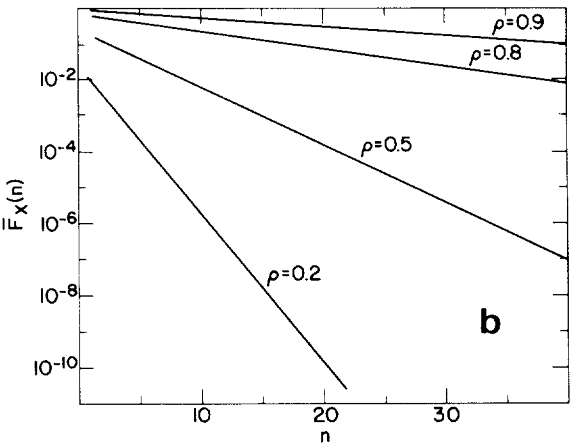


Fig. 4. Geometric arrivals and constant rate output. $\mathbf{P}\{x_k > n\}$ vs. n for various values of the traffic intensity: $\rho = 0.2, 0.5, 0.8$ and 0.9. (a) $\mu = 1$; (b) $\mu = 3$.

Table 1
z_0 for $\mu = 3$ and various values of r

ρ	r	z_0
0.2	$\frac{3}{8}$	2.56825
0.5	$\frac{3}{5}$	1.4463
0.8	$\frac{12}{17}$	1.12045
0.9	$\frac{27}{37}$	1.0546

Dividing by $(z_0 - 1)$ leads to

$$(z_0 - 1)\left(z_0^3 + \left(1 - \frac{1}{r}\right)z_0^2 + \left(1 - \frac{1}{r}\right)z_0 + \left(1 - \frac{1}{r}\right)\right) = 0. \qquad (4.21)$$

Since $(1 - 1/r)$ is negative, the polynomial of third degree has one real and two imaginary zeros. By means of numerical iteration the real root can be found rather fast. Table 1 shows the results for various r, where the traffic intensity is given by $\rho = 1/((1 - \mathrm{r})\mu)$.

4.3. Comparison with other results

In this section we compare our results with earlier work reported on the same subject. As mentioned earlier, Wyner [3] had formulated, prior to Kobayashi and Konheim [6], the buffer behavior problem as GI/G/1 system. He obtained the following results based upon quite elaborate arguments, which will not be reproduced here: If $\mathrm{E}[c_k] < 0$, then for some K_1, $K_2 > 0$ and z_0 as defined above

$$P_{\text{overflow}}(L) \approx K_1 z_0^{-L} \qquad (4.22)$$

and

$$p_L(L) \leqslant K_2 z_0^{-L}. \qquad (4.23)$$

Although our upper bound on $P_{\text{overflow}}(L)$ given by (3.40) does not directly imply the geometric form of (4.22), numerical evaluations of our $P_{\text{overflow}}(L)$ clearly indicate this property, as illustrated in Fig. 5(a) and (b). The agreement of our results with Wyner's results is of course not unexpected, but is a welcome confirmation of our bonding method.

Chu [7] discusses the buffer problems for the Poisson arrival and constant output case. He expressed the probability of the number of units at the end of the time slot in terms of the probability of the number presented at the beginning of the slot. Then a set of linear equations was solved numerically with the aid of the computer. The numerical results presented by Chu look similar to those of our study. Fig. 5 shows the upper bound (3.40) on P_{overflow} together with Chu's P_{overflow}. However, a complete agreement is not found. This discrepancy is a consequence of the difference in the model formulations. In Chu's model the incoming data stay in the buffer at least until the next time slot begins, whereas in our model the arriving data may instantly flow through as long as there is no blocking. Chu's formulation led to a recursive relation, which is equivalent to the im-

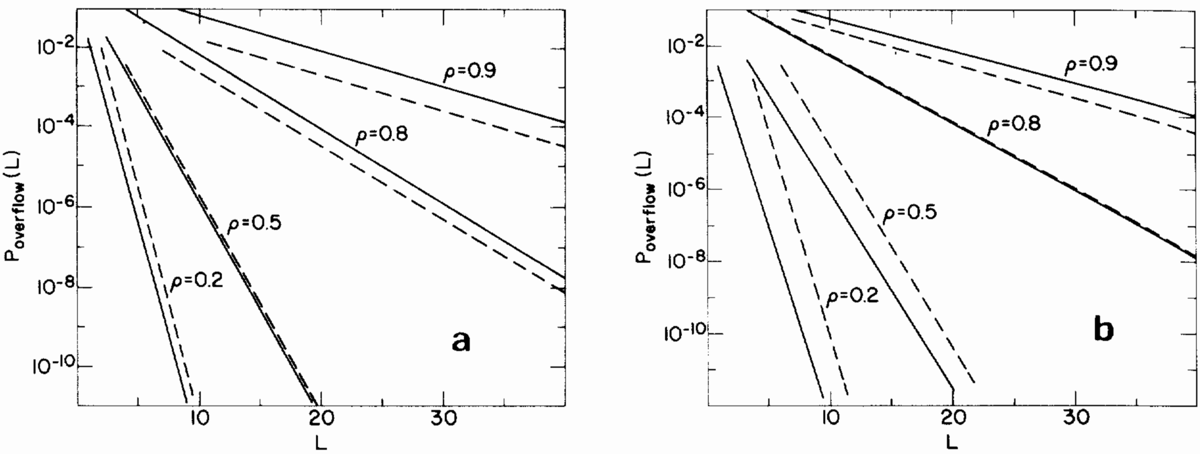


Fig. 5. Poisson arrivals and constant rate output. The upper bound on the overflow probability (———) compared with Chu's overflow probability (- - - - - -) for various values of the traffic intensity: $\rho = 0.2$, 0.5, 0.8 and 0.9. (a) $\mu = 1$; (b) $\mu = 3$.

bedded Markov chain analysis (see, e.g., [10,15]) developed for the M/G/1 queueing model. This difference between the two models is reflected in the numerical results plotted in Fig. 5 (for further details concerning the behaviour of the two models and the consequences thereof, see [14]). In a recent paper [16] Kobayashi shows that the characteristic equation that governs the asymptotic form of buffer overflow in the M/G/1 model is closely related to the characteristic equation (2.18) of the GI/G/1 model.

Appendix A

A.1. The characteristic root $z_0 > 1$

Let c be an integer-valued random variable with probability distribution $\{h_n\}$. By $H(z)$ we denote the generating function

$$H(z) = \mathrm{E}[z^c] = \sum_{n=-\infty}^{\infty} h_n z^n \tag{A1}$$

where $h_n = \mathbf{P}\{c = n\}$. Then under the assumptions

$$\mathrm{E}[c] < 0 \tag{A2}$$

and

$$\mathbf{P}\{c \geqslant 1\} > 0, \tag{A3}$$

the function $H(z)$ has exactly one real root $z_0 > 1$ that satisfies $H(z_0) = 1$.

A.2. Proof

Since $H(Z)$ contains terms z^{-n} $(n \geqslant 1)$, we have that

$$\lim_{z \to 0} H(z) = \infty. \tag{A4}$$

The so-called Markov inequality states that any non-negative random variables Y satisfies the relation

$$K\,\mathbf{P}\{Y \geqslant K\} \leqslant \mathrm{E}[Y] \quad \text{for any } K > 0. \tag{A5}$$

By setting $Y = z^c$ and $K = z$ we obtain

$$z\,\mathbf{P}\{z^c \geqslant z\} \leqslant \mathrm{E}[z^c] = H(z) \quad \text{for } z > 0. \tag{A6}$$

If we assume $z \geqslant 1$, then

$$z\,\mathbf{P}\{z^c \geqslant z\} = z\mathrm{P}\{c \geqslant 1\}. \tag{A7}$$

Then letting z approach infinity we obtain from (A3)

$$\lim_{z \to \infty} z\,\mathbf{P}\{c \geqslant 1\} = \infty. \tag{A8}$$

Since all the coefficients of $H(z)$ are non-negative, $H''(z) > 0$ for all real $z > 0$. Therefore $H(z)$ must have exactly one minimum z^* in the interval $(0, \infty)$. Because of the property (A2)

$$H'(1) = \mathrm{E}[c] < 0, \tag{A9}$$

the minimum point z^* must be located in the region $z^* > 1$, thus there must be a real number z_0 such that $z_0 > z^* > 1$ and $H(z_0) = 1$.

Appendix B

B.1. Wald's fundamental identity for random variables with discrete distributions

Let $\{x_k\}$ be i.i.d. integer-valued random variables and Y_i denote the sum of the first i elements of $\{x_k\}$:

$$Y_i = x_1 + x_2 + \cdots + x_i. \tag{B1}$$

Let j be the stopping time corresponding to the interval $[b, a]$, i.e.,

$$j = \min\{i:\ Y_i < b \text{ or } Y_i > a\}. \tag{B2}$$

Assume that the random variable x satisfies the following conditions:

(1) The mean and variance of x both exist and the variance is strictly positive.

(2) $\mathbf{P}\{x > 0\} > 0$ and $\mathbf{P}\{x < 0\} > 0$. (B3)

(3) For some $r > 1$ the expectation

$$\mathrm{E}[z^x] = G(z) \tag{B4}$$

exists for z such that $0 < z < r$.

Then, the following equality holds for all z such that $G(z) \geqslant 1$:

$$\mathrm{E}\left[z^{Y_j} G(z)^{-j}\right] = 1. \tag{B5}$$

B.2. Proof

Let J be a positive integer. Then for any stopping time $j \geqslant 1$

$$\begin{aligned}\mathrm{E}[z^{Y_j + (Y_J - Y_j)}] &= \mathrm{E}[z^{Y_J}] \\ &= \mathrm{E}[z^{x_1 + x_2 + \cdots + x_J}] \\ &= G(z)^J. \end{aligned}\tag{B6}$$

Let us define

$$P_J = \mathbf{P}\{j \leqslant J\}, \tag{B7}$$

$$E_J[u] = \mathrm{E}[u \mid j \leqslant J] \tag{B8}$$

and

$$\bar{E}_J[\mathrm{u}] = \mathrm{E}[u \mid j > J]. \tag{B9}$$

Then, we obtain from (B6)

$$\begin{aligned}\mathrm{E}[z^{Y_J}] &= P_J E_J[z^{Y_j + (Y_J - Y_j)}] + (1 - P_J)\bar{E}_J[z^{Y_J}] \\ &= G(z)^J. \end{aligned}\tag{B10}$$

The fact that the stopping time j implies that the variables $x_1, x_2, \ldots, x_j$ are no longer independent, since they must satisfy $Y_j > a$ or $Y_j < b$. If $j \leqslant J$, then $Y_J - Y_j = x_{j+1} + x_{j+2} + \cdots + x_J$, is independent from $Y_j = x_1 + x_2 + \cdots + x_j$, so we can write

$$E_J[z^{Y_j + (Y_J - Y_j)}] = E_J\left[z^{Y_j} G(z)^{J-j}\right]. \tag{B11}$$

Therefore (B10) becomes

$$G(z)^J = P_J E_J\left[z^{Y_j} G(z)^{J-j}\right] + (1 - P_J)\bar{E}_J[z^{Y_J}]. \tag{B12}$$

Dividing both sides by $G(z)^J$ leads to

$$P_J E_J\left[z^{Y_j} G(z)^{-j}\right] + (1 - P_J)\frac{\bar{E}_J[z^{Y_J}]}{G(z)^J} = 1. \tag{B13}$$

Now, obviously,

$$\lim_{J \to \infty} P_J = \lim_{J \to \infty} \mathrm{P}\{j \leqslant J\} = 1. \tag{B14}$$

Hence,

$$\lim_{J \to \infty} (1 - P_J) = 0. \tag{B15}$$

Since $j > J$ only if $b < Y_J < a$, we find that for $z \geqslant 1$

$$z^b < z^{Y_J} < z^a. \tag{B16}$$

Hence, $\bar{E}_J[z^{Y_J}]$ is finite for any J, and using the assumption that $G(z) \geqslant 1$, we obtain

$$\lim_{J \to \infty} (1 - P_J)\frac{\bar{E}_J[z^{Y_J}]}{G(z)^J} = 0. \tag{B17}$$

Furthermore,

$$\lim_{J \to \infty} P_J E_J\left[z^{Y_j} G(z)^{-j}\right] = \mathrm{E}\left[z^{Y_j} G(z)^{-j}\right]. \tag{B18}$$

Thus, from (B13), we obtain the desired identity (B5).

References

[1] H. Kobayashi, Bounds on buffer overflow probabilities in communication systems, Internat. Symp. on Information Theory, Grignano, 1979.

[2] M. Schwartz, Computer Communication Network Design and Analysis (Prentice-Hall, Englewood Cliffs, NJ, 1977).

[3] A.D. Wyner, On the probability of buffer overflow under an arbitrary bounded input–output distribution, SIAM J. Appl. Math. (4) (1974) 544–570.

[4] J.F.C. Kingman, Inequalities in the theory of queues, J. Roy. Statist. Soc. Ser. B 32 (1970) 102–110.

[5] S.M. Ross, Bounds on the delay distribution in GI/G/1 queues, J. Appl. Probab. 2 (1974) 417–421.

[6] H. Kobayashi and A.G. Konheim, Queueing models for

computer communication system analysis, IEEE Trans. Comm. Com-25 (1) (1977) 2–29.

[7] W.W. Chu, Buffer behavior for Poisson arrivals and multiple synchronous constant outputs, IEEE Trans. Comput. C-19 (6) (1970) 530–534.

[8] D.V. Lindley, The theory of queues with a single server, Proc. Cambridge Phil. Soc. 49 (1952) 277–289.

[9] H. Kobayashi, Bounds for the waiting time in queueing systems, in: Computer Architecture and Networks (North-Holland, New York, 1974) pp. 263–274.

[10] L. Kleinrock, Queueing Systems Vol. I: Theory, Vol II: Computer Applications (Wiley, New York, 1974, 1975).

[11] A. Wald, Sequential Analysis (Wiley, New York, 1947).

[12] M. Fisz, Probability Theory and Mathematical Statistics (Wiley, New York, 3rd. ed., 1963).

[13] R.E. Barlow and F. Proschan, Mathematical Theory of Reliability (Wiley, New York, 1967).

[14] K. Geihs, Grenzen für die Pufferüberlaufwahrscheinlichkeit, Diplom Thesis, Fachbereich Informatik, Technische Hochschule Darmstadt, 1980.

[15] H. Kobayashi, Modeling and Analysis (Addison-Wesley, Reading, MA, 1978).

[16] H. Kobayashi, Discrete-time queueing systems, IBM Res. Rept. RC8817, 1981; in: Probability in Computer Science (Academic Press, New York) to appear.