W. M. Chow
L. Woo

Buffer Performance Analysis of Communication Processors During Slowdown at Network Control

Abstract: An approximate model based on renewal theory has been developed for the performance analysis of communication processors during Network Control Program (NCP) slowdown. Performance values, which include cycle length, buffer utilization, and message loss, were computed as functions of traffic loads and user-assigned threshold values of free buffers were used for slowdown control. Comparison of results obtained by the approximate method with simulated results shows a high degree of accuracy for the approximation.

Introduction

Teleprocessing in data communication systems is a rapidly growing part of the data processing industry. The complexity of a teleprocessing network increases with the diversity of teleprocessing products, communication facilities, transmission control units, and other equipment. In earlier installations, primary control of the teleprocessing network was performed by the central processing unit (CPU), with an access method administering the flow of data from the stations to the CPU and vice versa. Because of this additional requirement for message handling, the performance of the CPU as a resource for application processing was degraded. With the introduction of communication processors such as the IBM 3704 or 3705 [1], this problem has been partially alleviated. Many message control functions have been transferred from the CPU to the communication processors.

Within the environment of the IBM Virtual Telecommunication Access Method and the Network Control Program (VTAM and NCP), the IBM 3704 or 3705 interacts with the communication scanner and the channel adapter in controlling flow of data through the network. In order to achieve high productivity, the performance of the 3704 or 3705 with the NCP software must be evaluated. Storage and performance estimates for this purpose are provided in published material [2, 3].

One of the major concerns in performance is the storage buffer pool for temporary storage of message data. Both inbound and outbound messages must reside in the buffer pool during the processing time of the NCP which includes error checking, block handling, and message routing. Buffer pool estimates are usually made based on the traffic requirements of the network [2]. Under normal operating conditions, the buffer pool will be adequate for handling message traffic. During periods of peak load, however, sporadic bursts of messages may occasionally cause the buffer pool to overflow. Since many of the messages are sequenced, loss or retransmission of too many messages could severely degrade NCP performance and lead to a possible deadlock [2, 3].

To partially relieve this situation, a slowdown algorithm has been implemented in the NCP software [4]. The algorithm provides two threshold values of free buffer space for controlling message arrival. When the free space in the buffer pool is reduced to the lower threshold value, pooling for terminal messages to the NCP is temporarily suspended and the subsequent depletion of messages in the buffer pool will free the buffer units. Arrival traffic resumes when the free space reaches the higher threshold value. Since the long-term overall arrival rate averages out to the normal operating condition, the system will eventually settle down after the sporadic bursts.

In this paper, we study the mechanism of the slowdown process in detail and analyze system performance in terms of buffer utilization and message loss. The following sections describe the slowdown process, give a mathematical method of analysis based on renewal theory, and present some illustrative examples of computation.

Slowdown process in the NCP—traffic rates and buffer usage

Although the sequence of events that takes place in the NCP during the slowdown process is difficult to describe

exactly, an approximate model may be used to depict the process in terms of repeated cycles (Fig. 1). Each cycle consists of five consecutive phases which are described as follows:

As buffers, or buffer units, from the buffer pool are used, the number of free buffer units, or the total free area, is decreased. At times of sporadic heavy load, the number of free buffer units drops down to or below the lower threshold value L. Consequently system slowdown takes effect, occurring at t_3 in Fig. 1.

The slowdown period, from t_3 to τ_{i+1} in Fig. 1, may be divided into two phases. During phase one (t_3 to t_4), the NCP stops polling messages and also issues a halt of traffic request to the VTAM host. Messages continue to arrive at a slower rate until residual messages in the network are cleared. The length of this phase (T_4) corresponds roughly to the transmission time of all messages from some station during a single poll. Since messages continue to arrive while buffer capacity is limited, there is a possibility of message loss. The second phase, from t_4 to τ_{i+1}, represents a period of depletion, during which arrival traffic is stopped and buffers are returned to the free space. System slowdown ends and the recovery period commences when the number of free buffer units reaches the upper threshold value M at time τ_{i+1}.

The recovery period, from τ_i to t_2, may again be described in two phases. During phase one (τ_i to t_1), although the NCP starts polling, there is a delay in the arrival of messages. The length of this phase (T_1) is roughly equal to the transmission time of a poll message. The second phase, from t_1 to t_2, experiences an abnormally high rate of traffic because of the accumulation of messages at stations during the slowdown process. The length of this phase (T_2) and the traffic rate may be estimated from the length of the slowdown period and the amount of suppressed traffic. At time t_2, the normal traffic period resumes. This continues until the next point (t_3) at which the lower threshold L is again reached. This period is shown as T_3.

Total buffer pool size is usually determined from normal traffic conditions [2]. With a fixed buffer pool size, the assigned values of the thresholds L and M affect buffer utilization and message loss. In general, the allowable percentage loss of messages is held below a certain value, while utilization of buffers may be optimized under given rates of traffic and processing.

The following sections present an analytical method for evaluating buffer utilization and loss of messages.

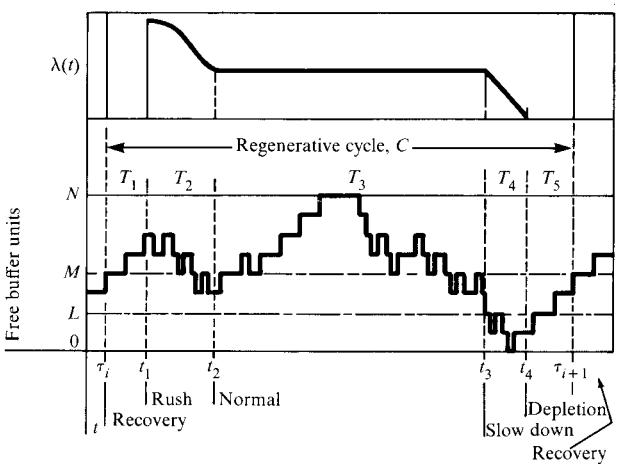

Figure 1 Number of free buffer units $k(t)$.

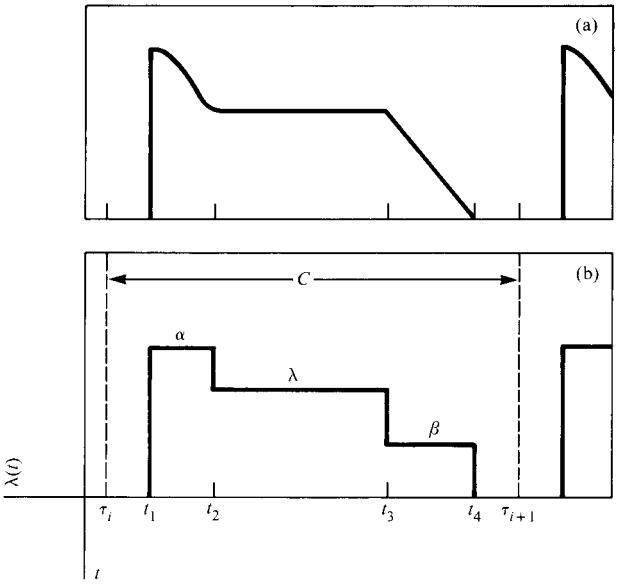

Figure 2 Message arrival intensity $\lambda(t)$.

Assumptions

The arrival stream is assumed to be a Poisson process with a rate $\lambda(t)$. Figure 2(a) depicts a typical arrival rate $\lambda(t)$ along the time axis. In this analysis, it is assumed that $\lambda(t)$ is a jump process [see Fig. 2(b)] such that

$$\lambda(t) = \begin{cases} 0 \text{ during the recovery and depletion periods,} \\ \alpha \text{ during the rush period,} \\ \lambda \text{ during the normal period, and} \\ \beta \text{ during the slowdown period,} \end{cases}$$

and $\beta < \lambda < \alpha$.

The message processing rate μ remains constant at all times and corresponds to the average rate of release of buffer units at the NCP.

The random process $K(t)$, which represents the number of *free* buffer units at time t, is shown in Fig. 1. For a stable system the process $K(t)$ must not drop down to the slowdown threshold L during the rush period, i.e., $P[K(t) > L]$ is almost equal to unity for $t_1 \leq t < t_2$. Similarly, it is required that $P[K(t) < M]$ be close to unity for $t_3 \leq t < t_4$.

The periods T_1, T_2 and T_4 are constant.

All messages are of the same length as the buffer unit size.

Analysis

The expected number of lost messages (in terms of buffer units) during a slowdown period and the extent of buffer utilization can be derived based on the behavior of the process $K(t)$. Two cases are investigated, the first for exponentially distributed processing time, and the second for constant processing time. In the NCP problem, the constant case is closer to reality.

It is well known [5] that expectations derived from a regenerative process can be obtained by looking at one single regenerative cycle, and this approach is adopted for our studies. Let $\{\tau_i, i = 1, 2, \cdots\}$ represent a sequence of time epochs such that a recovery period begins at each τ_i. Since the arrival of message units is assumed to be a Poisson process and a recovery period is always initiated by having $K(\tau_i^+) = M$, the upper threshold value, it is clear that $\{\tau_i\}$ forms a regenerative process with a cycle length $C = \tau_{i+1} - \tau_i$ (see Fig. 1).

For convenience, it is further assumed that $t = 0$ is a regenerative point. Both the cumulative number of arrivals $A(t)$ and the cumulative number of departures $D(t)$ have a zero value at $t = 0$.

Buffer utilization is then defined by

$$u = 1 - \frac{\mathrm{E}\left[\int_0^C K(t)\, dt\right]}{N \cdot \mathrm{E}[C]}, \tag{1}$$

where N is the buffer pool size. The numerator in Eq. (1) represents expected total free area (buffer-time units) during C, which will be evaluated by separately considering the periods $\{T_i, i = 1, \cdots, 5\}$. Let

$$t_i = \sum_{j=1}^{i} T_j, \qquad i = 1, 2, \cdots, 5.$$

The free area during each period is given by

$$A_i = \int_{t_{i-1}}^{t_i} K(t)\, dt, \qquad i = 1, 2, \cdots, 5, \tag{2}$$

and

$$0 = t_0 < t_1 < \cdots < t_5 = C.$$

Then the total free area

$$A = \int_0^C K(t)\, dt = \sum_{i=1}^{5} A_i.$$

The expected number of lost messages during C is given by

$$V = \mathrm{E}[A(C) - D(C)].$$

For a stable system, message loss can occur only during the slowdown period T_4. Since $K(t_3) = L$, the lower threshold value,

$$\begin{aligned} V &= L + \mathrm{E}[A(t_4) - A(t_3)] - \mathrm{E}[D(t_4) - D(t_3)] \\ &\quad - \mathrm{E}[K(t_4)] \\ &= L + \beta T_4 - \mathrm{E}[D(t_4) - D(t_3)] - \mathrm{E}[K(t_4)]. \end{aligned} \tag{3}$$

• *Case 1: Exponential Processing Time*

When exponential processing time distribution is considered, the process $K(t)$ can be characterized by the queue size in an M/M/1 queuing system with a waiting room of capacity N, an *arrival* rate $\mu(t)$, and a *service* rate $\lambda(t)$, i.e., interarrival time and service time in the queuing system of the process $K(t)$ are identical to message processing time and interarrival time, respectively.

To avoid any ambiguity, the process $K(t)$ will be considered in terms of queue size in the following analysis, and all terms such as arrival, departure, etc. are referred to this queuing system.

Expected number of lost messages

Since it is unlikely that $K(t)$ will attain the value N during the slowdown period, $D(t_4) - D(t_3)$ can be viewed as a Poisson process with a rate μ(the exponential rate of arrivals to the queuing system). Equation (3) can then written as

$$V = L + \beta \cdot T_4 - \mu \cdot T_4 - \mathrm{E}[K(t_4)]. \tag{4}$$

It is known [6] that the transient behavior of the queue size $K(t)$ in an M/M/1 system is characterized by

$$\begin{aligned} \mathrm{P}_{ij}(S) &= P[K(S) = j | K(0) = i] \\ &= r_{j-i}(S) + \rho^{-i-1} r_{j+i+1}(S) + (1-\rho)\rho^j R_{-j-i-2}, \end{aligned} \tag{5}$$

where

$$\rho = \mu/\beta,$$

$$r_j(S) = \sum_{n=0}^{\infty} e^{-\mu S} \frac{(\mu S)^{n+j}}{(n+j)!} e^{-\beta S} \frac{(\beta S)^n}{n!}, \text{ and}$$

$$R_j(S) = \sum_{i=-\infty}^{j} r_i(S).$$

Based upon this relation, it can be shown that

$$\mathrm{E}[K(S)|K(0)=i] = i + (\mu-\beta)S + \frac{\rho}{1-\rho}$$
$$\times \exp\{\beta(\rho-1)S + \mu(\rho^{-1}-1)S\}$$
$$-\frac{\rho^{-i}}{1-\rho} + \frac{\rho}{1-\rho}\sum_{j=0}^{i}(\rho^{-i-1}-\rho^{-j})\, r_j(S)$$
$$+\sum_{j=1}^{i-1}(i-j)\,\rho^{-j} r_j(S). \qquad (6)$$

Since $P[K(t_3)=L]=1$,

$$\mathrm{E}[K(t_4)] = \mathrm{E}[K(t_3+T_4)|K(t_3)=L],$$

which is given by Eq. (6) with S replaced by T_4 and i by L. Thus, the expected number of lost messages V can be determined by Eqs. (4) and (6).

Expected free area during recovery period T_1

The recovery period starts at $t=0$ with $K(0^+)=M$. Since $\lambda(t)=0$, for $t\in[0,t_1)$, the queue size $K(t)$ grows monotonically at a rate μ. For a given $K(t_1)$, it is well known that the arrival epochs are uniformly distributed over the interval $[0,t_1)$; hence

$$\mathrm{E}[A_1|K(t_1)] = K(0)T_1 + \tfrac{1}{2}\mathrm{E}[K(t_1)-K(0)]T_1,$$

and

$$\mathrm{E}[A_1] = M\cdot T_1 + \tfrac{1}{2}\mu\cdot T_1^2. \qquad (7)$$

Expected free area during rush period T_2

During this period, the server resumes his service with a rather high rate $\alpha>\mu$. In order to evaluate

$$\mathrm{E}[A_2] = \int_{t_1}^{t_2}\mathrm{E}[K(t)]\,dt,$$

one may first compute the conditional expectation $\mathrm{E}[K(t)|K(t_1)]$ by employing Eq. (6) to obtain $\mathrm{E}[K(t)]$. This approach, however, requires a tedious computational effort. A simple approximation of sufficient accuracy is recommended in the following:

By assuming that $P[L<K(t)<N, t\in[t_1,t_2)]\approx 1$, it follows that

$$K(t)\approx K(t_1) + [D(t)-D(t_1)] - [A(t)-A(t_1)]. \qquad (8)$$

Then

$$\mathrm{E}[K(t)]\approx M + \mu T_1 - (\mu-\alpha)(t-t_1),$$

and

$$\mathrm{E}[A_2]\approx \int_0^{T_2}[M+\mu T_1-(\mu-\alpha)t]\,dt$$
$$= (M+\mu T_1)T_2 + (\alpha-\mu)\,T_2^2/2. \qquad (9)$$

The accuracy of Eq. (9) depends upon parameters α, μ and T_2. In general, α should be large compared with μ so that $K(t)<N$. On the other hand, T_2 must be small enough that $K(t)>L$.

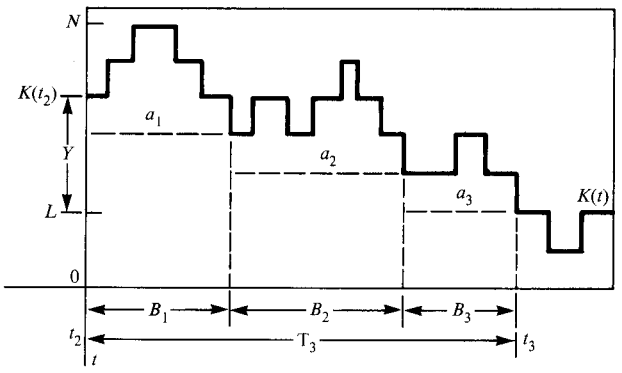

Figure 3 Busy periods and areas under $k(t)$ during normal period T_3, when $k(t_2)-L=3$.

Expected free area during normal period T_3 and expected length of the normal period

After a rush period, the service rate is back to normal; i.e., $\lambda(t)=\lambda$. The termination of the normal period is at the instant when $K(t)=L$, and therefore T_3 is a random variable. In this situation, one should evaluate not only $\mathrm{E}[A_3]$ but also $\mathrm{E}[T_3]$. This can be done by introducing the concept of a busy period.

A busy period is initiated by an arrival that finds an idle server, and terminated when the system first becomes idle. Thus, a busy period is identically distributed as an interval $[x,y)$, such that $K(y)=K(x)-1$ and $K(z)>K(y)$, $\forall\, z\in[x,y)$. If $K(t_2)=n+L$, then the period T_3 contains n distinct busy periods $(B_1,\cdots,B_n)$ and B_i is defined by the busy period of an M/M/1 queue with a finite waiting room of capacity $N-K(t_2)+i$. (See Fig. 3.) It can be shown that

$$\mathrm{E}[B_i] = \frac{1}{\lambda-\mu}(1-\rho^Q), \qquad (10)$$

where $\rho=\mu/\lambda$, and $Q=N-K(t_2)+i$. Thus,

$$\mathrm{E}[T_3|K(t_2)=k] = \sum_{Q=N-k+1}^{N-L}\frac{1}{\lambda-\mu}(1-\rho^Q)$$
$$= \frac{k-L}{\lambda-\mu} - \rho^{N-k+1}\frac{1-\rho^{k-L}}{1-\rho}, \text{ and}$$
$$\mathrm{E}[T_3] = \frac{\mathrm{E}[K(t_2)]-L}{\lambda-\mu} - \frac{\rho^{N+1}\mathrm{E}[\rho^{-K(t_2)}]-\rho^{N-L+1}}{1-\rho}. \qquad (11)$$

Let a_i be the expected area under the curve of the queue size in an M/M/1 queue during B_i, and $Y=K(t_2)-L$. Given Y, the conditional expected free area (see Fig. 3) is

$$\mathrm{E}[A_3|Y] = \sum_{i=1}^{Y}\{a_i + (L+Y-i)\mathrm{E}[B_i]\}. \qquad (12)$$

Note that when the waiting room capacity is $Q = N - K(t_2) + i$, the expected queue size is

$$L(Q) = \lim_{t\to\infty} \int_0^t N(S)\, dS/t$$

$$= \frac{\mathrm{E}\left[\int_0^{B_i} N(S)\, dS\right]}{\mathrm{E}[B_i + I]},$$

where $N(S)$ is the queue size at time S and I is the idle period. By definition,

$$a_i = \mathrm{E}\left[\int_0^{B_i} N(t)\, dt\right]$$

$$= L(Q)\ \{\mathrm{E}[B_i] + 1/\mu\}. \tag{13}$$

Furthermore, it is easy to show that

$$L(Q) = \frac{\rho}{1-\rho} - \frac{(Q+1)\,\rho^{Q+1}}{1-\rho^{Q+1}}. \tag{14}$$

By using Eqs. (10), (12), (13), and (14), it follows that

$$\mathrm{E}[A_3] = \frac{1}{\lambda-\mu}\left\{\frac{\mathrm{E}[Y^2]}{2} + \left(\frac{1}{1-\rho} - \frac{1}{2}\right)\mathrm{E}[Y] + \left(n + \frac{1}{1-\rho}\right)\frac{\rho^{n+1}}{1-\rho}\left(1 - \mathrm{E}[\rho^{-Y}]\right)\right\}, \tag{15}$$

where $n = N - L - 1$ and $Y = K(t_2) - L$.

By using Eq. (8), one can estimate the expectations of Y in Eq. (15) as follows:

$$Y \approx M + [D(t_2) - D(0)] - [A(t_2) - A(0)] - L$$

$$= M + D(t_2) - [A(t_2) - A(t_1)] - L.$$

Since $D(t_2)$ and $A(t_2) - A(t_1)$ have approximately independent Poisson distributions with means $\lambda(T_1 + T_2)$ and αT_2, respectively,

$$\mathrm{E}[Y] \approx M - L + \mu T_1 + (\mu - \alpha)T_2, \tag{16}$$

$$\mathrm{E}[Y^2] \approx \mu(T_1 + T_2) + \alpha T_2 + (\mathrm{E}[Y])^2, \text{ and} \tag{17}$$

$$\mathrm{E}[\rho^{-Y}] = \rho^{L-M} \exp\{[\lambda(T_1 + T_2) - \alpha T_2]\ (1-\rho)\}. \tag{18}$$

By substituting Eqs. (16)–(18) into Eq. (15), $\mathrm{E}[A_3]$ is obtained.

Expected free area during slowdown period T_4

The slowdown period starts when $K(t)$ is reduced to L. Then, $\lambda(t) = \beta < \mu$. Clearly

$$\mathrm{E}[A_4] = \mathrm{E}\left[\int_{t_3}^{t_4} K(t)\, dt\right]$$

$$= \int_0^{T_4} \sum_{j=0}^{\infty} j P_{Lj}(t)\, dt,$$

where $P_{Lj}(t)$ is defined by Eq. (5).

Since L is so chosen that the possibility of message loss is negligible, a simple approximation for $\mathrm{E}[A_4]$ can be derived as follows. For

$$K(t) \approx L + [D(t) - D(t_3)] - [A(t) - A(t_3)], \qquad t \in [t_3, t_4),$$

$$\mathrm{E}[K(t)] = L - (\beta - \mu)(t - t_3). \tag{19}$$

Therefore,

$$\mathrm{E}[A_4] = LT_4 - \tfrac{1}{2}(\beta - \mu)\ T_4^2. \tag{20}$$

Expected area during the depletion period T_5 and expected length of the depletion period

During this period, no customer will be served (i.e., there are no incoming messages), hence $\lambda(t) = 0$. The period is terminated when queue size is equal to M. It follows immediately that $K(t) - K(t_4)$ is a Poisson process with a rate μ and $K(\tau_i + C) = K(\tau_{i+1}) = M$.

$$\mathrm{E}[T_5 | K(t_4)] = \frac{1}{\mu}\,[M - K(t_4)],$$

and

$$\mathrm{E}[A_5 | K(t_4)] = \tfrac{1}{2}\mathrm{E}[T_5 | K(t_4)]\ (M - K(t_4) - 1) + \mathrm{E}[T_5 | K(t_4)]\ K(t_4).$$

Taking expectations on both sides of the equations and using Eq. (19), we have

$$\mathrm{E}[T_5] = \frac{1}{\mu}\,[M - L + (\beta - \mu)\ T_4], \text{ and} \tag{21}$$

$$\mathrm{E}[A_5] = \frac{1}{2\mu}\{M(M-1) + L - 2\beta \cdot T_4 + [L - (\beta - \mu)\ T_4]^2\}. \tag{22}$$

• *Case 2: Constant Processing Time*

When the processing time is *constant*, say d, the random process $K(t)$ or its complement $N(t) = N - K(t)$ no longer possesses the memoryless property. This problem can be handled by means of the imbedded Markov chain as in the analysis of the M/D/1 queue with finite waiting room of capacity N. The regenerative cycle has the same components as in the exponential case. Using the same approach as before, the separate analysis of the five periods (T_i, $i = 1, 2, \cdots, 5$) may be carried out. The complication of this analysis lies in the fact that, with the exception of T_1, each of the periods may not initiate at the instant of service completion. In the case of T_2 and T_3, the remaining service times of the first message, S_2 and S_3, are fixed values. For the periods T_4 and T_5, the remaining service times of the first message, S_4 and S_5, are random variables. In the appendix, we outline a method of determining the distribution of S_4. Knowing the distribution of S_4, we may easily find the distribution of S_5 since T_4 is constant.

For practical problems, the number of messages served during each period is so large that the effect of first service time on utilization is rather small and can usually be neglected.

The determination of free area A_i or its complement $NT_i - A_i$ is straightforward for the periods T_1 and T_5 since $\lambda(t) = 0$ for these periods. For T_2 and T_4, the computation is more complicated. One may evaluate the process $N(t)$ through a few consecutive processing intervals of length d each.

For the normal period T_3, which takes up a large percentage of the cycle time, we shall describe in detail the method of evaluating 1) the expected length of the period $E[T_3]$, and 2) the expected used buffer area $E[N \cdot T_3 - A_3]$.

Since the normal period may not start at a service completion, we discuss the first message service time separately from the subsequent messages.

Distribution of $N(t)$ at the end of first message service time, $N(t_2 + S_3)$.

Note that

$$N(t) \approx N - M + A(t) - D(t)$$
$$= N - M + A(t) - h(t/d), \qquad t \geq 0, \tag{23}$$

where $h(X)$ is the integer part of X. Since the Poisson arrival stream has a rate

$$\lambda(t) = \begin{cases} 0, & 0 \leq t < t_1; \\ \alpha, & t_1 \leq t < t_2; \\ \lambda, & t_2 \leq t < t_2 + S_3. \end{cases}$$

$$P[N(t_2 + S_3) = n] = \sum_{i=0}^{R} e^{-\alpha T_2} \left[\frac{(\alpha T_2)^i}{i!}\right] e^{-\lambda S_3} \left[\frac{(\lambda S_3)^{R-i}}{(R-i)!}\right], \tag{24}$$

where $R = n - N + M + h\ (t_2 + S_3/d)$, and S_3 is the fraction part of t_2/d.

Expected length of the normal period T_3

The length of a normal period can be interpreted as the first passage time of the event $\{N(t) = N - L\}$. Let

$$B_n = \min\ \{t | N(t) = N - L,\ N(0) = n,\ N(0^-) = n + 1\},$$

and

$$b_n = E[B_n].$$

Clearly,

$$E[T_3] = S_3 + \sum_n b_n \cdot P[N(t_2 + S_3) = n]. \tag{25}$$

It is then required to evaluate b_n. Define a Markov chain imbedded at each departure epoch, and let

$$q_j = P[\text{there are } j \text{ arrivals during time interval } d]$$
$$= e^{-\lambda d} \frac{(\lambda d)^j}{j!}. \tag{26}$$

Suppose that a departure occurs at t and $N(t^+) = n$. If $n = 0$, the next transition must be an arrival. In the case in which $n > 0$ and there are k arrivals during the next service time d, then $N(t + d) = (n - 1) + k$. For $k \geq (N - L) - n$,

$$E[B_n | k] = \frac{(N - L) - n}{k + 1} d.$$

Consequently, a set of linear equations can be defined as

$$\begin{cases} b_n = \sum_{j=0}^{N-L-n-1} (d + b_{n-1+j})\ q_j \\ \qquad + \sum_{j=N-L-n}^{\infty} d \cdot \frac{N - L - n}{j + 1} \cdot q_j, \\ \qquad n = 1, 2, \cdots, N - L - 2; \\ b_0 = \frac{1}{\lambda} + b_1. \end{cases} \tag{27}$$

In solving for $\{b_n\}$, the value of $E[T_3]$ can be determined by using Eqs. (24) and (25).

Expected used buffer area $E[N \cdot T_3 - A_3]$ during the normal period

Define

$$a_n = E\left[\int_0^{B_n} N(t)\ dt | N(0) = n,\ N(0^-) = n + 1\right].$$

If there are k arrivals during an interdeparture time d, then the expected area a_n will be

$$\begin{cases} (nd + \frac{k}{2} d) + a_{n-k+1}, \\ \qquad \text{if } k = 0, 1, \cdots, N - L - n - 1; \\ \left(n + \frac{N - L - n - 1}{2}\right) d \frac{N - L - n}{k + 1}, \\ \qquad \text{if } k \geq N - L - n. \end{cases}$$

It follows that

$$\begin{cases} a_n = \sum_{K=0}^{N-L-n-1} \left(\frac{K + 2n}{2} d + a_{n+K-1}\right) q_K \\ \qquad + \sum_{K=N-L-n}^{\infty} \left(\frac{N - L + n - 1}{2}\right)\left(\frac{N - L - n}{K + 1}\right) d q_K, \\ \qquad n = 1, \cdots, N - L - 2; \\ a_0 = a_1. \end{cases} \tag{28}$$

The conditional expectation of the used buffer area during S_3, given that $N(t_2 + S_3) = n$, is

Table 1 Computation of simulation and analytic results.

Given Data:

$\alpha = 44.8$	Message size 256 bytes
$\lambda = 38.4$	$L = 8$ message units
$\beta = 32$	$M = 32$ message units
$\mu = 36$	$N = 64$ message units

	Constant Proc. Time		*Exponential Proc. Time*	
	Simulation	*Analytic*	*Simulation*	*Analytic*
T_1	0.1		0.1	
T_2	0.5		0.5	
T_3	9.67	9.66	9.56	9.08
T_4	1		1	
T_5	0.532	0.556	0.544	0.556
(util) T_1	0.479		0.470	0.472
(util) T_2	0.483		0.479	0.478
(util) T_3	0.562	0.569	0.496	0.490
(util) T_4	0.838		0.837	0.844
(util) T_5	0.670		0.683	0.691
C	11.9	11.81	11.8	11.24
Utilization	0.588	0.594	0.535	0.531
Message Loss (%)	2.5×10^{-4}		1.5×10^{-3}	8.4×10^{-4}

$$W_n = \sum_{i=0}^{R} \left[\frac{S_3 \cdot (R-i)}{2} + i \cdot S_3\right] e^{-\alpha T_2} \frac{(\alpha T_2)^i}{i!} \times e^{-\lambda S_3} \frac{(\lambda S_3)^{R-i}}{R-i} \left(P[N(t_2 + S_2) = n]\right)^{-1}, \quad (29)$$

where $R = n - N + M + h\,(t_2 + S_3/d)$. From Eqs. (24), (28), and (29), the value of

$$\mathrm{E}[N \cdot T_3 - A_3] = \sum_n (W_n + a_n)\, P[N(t_2 + S_3) = n] \quad (30)$$

can be determined.

Numerical results and discussion of results

A sample problem illustrating the 3705 NCP process was chosen with the following data:

A fixed size message length of 256 bytes is assumed. The buffer pool N is 16K bytes, or the equivalent of 64 message units. The NCP processing rate is estimated to be 36 message units per second.

Three combinations of threshold values (M, L) in terms of message units are selected for study. They are (32, 8), (32, 16) and (16, 8).

Arrival rate λ of traffic to the NCP during the normal phase T_3 varies from 38 to 46 message units per second. The corresponding rush phase rate α during T_2 varies from 44 to 52 message units per second.

The lengths of the recovery phase T_1, the rush phase T_2, and the slowdown phase T_4 are estimated to be 0.1, 0.5, and 1.0 respectively.

The cases for both constant and exponential processing time are investigated for the normal phase T_3. For the other phases, computations are based on exponential processing time. Results are verified through a detailed simulation using APLOMB [7]. Simulation results confirm our assumption that buffer utilization during T_1, T_2, T_4, and T_5 varies slightly with respect to the processing time distribution (exponential or constant). Total buffer utilization, however, differs significantly for the two cases.

From simulation results, it is observed that when $(M, L) = (16, 8)$, the values of M and L are so close together that a stable system cannot be realized. The process simply oscillates between slowdown and rush phases.

- *Computation of simulation and analytic results*

Computation results and the chosen conditions on which the computations were based are given in Table 1.

- *Cycle length, buffer utilization, and fractional loss of message*

Figures 4, 5, and 6 give the plots of cycle length, overall buffer utilization and fractional message loss against traffic arrival rate λ.

The $(M, L) = (32, 8)$ case gives a higher utilization than does the $(M, L) = (32, 16)$ case, but the fractional message loss is also higher.

- *Discussion of results*

When the buffer pool size N is given, the designer may choose the smallest value of L such that the fractional message loss can be kept within an allowable limit. He may then select the value of M which gives the best utilization. However, buffer utilization governs the mean

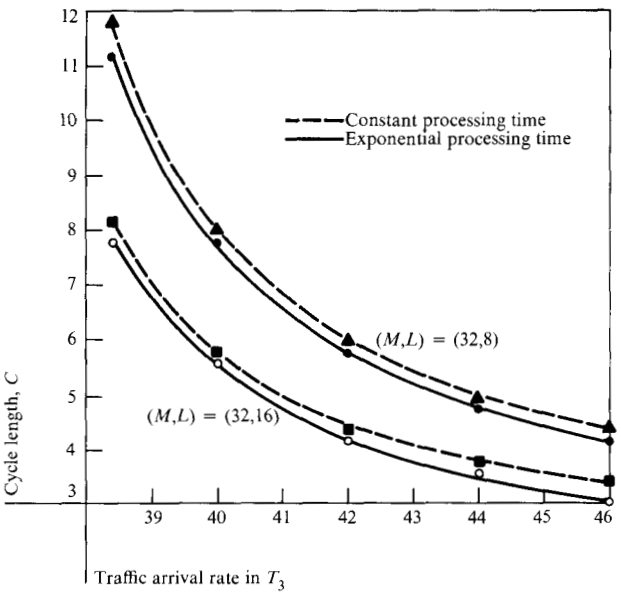

Figure 4 Expected cycle length vs traffic arrival rate in T_3.

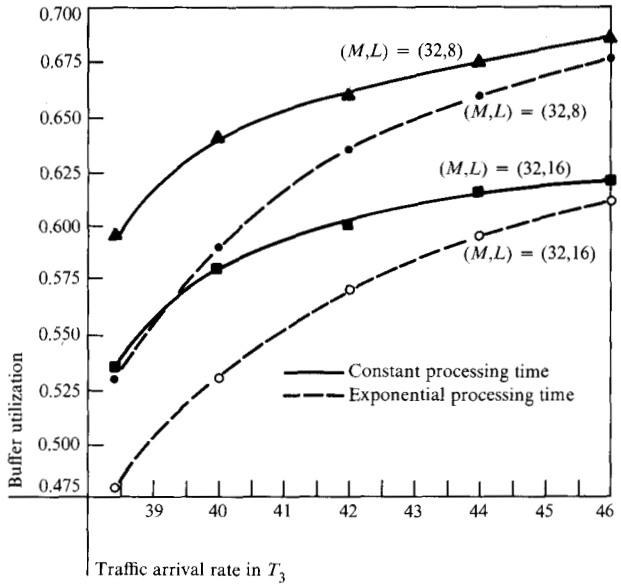

Figure 5 Buffer utilization vs traffic arrival rate in T_3.

effective traffic arrival rate, which is always less than λ. The net result of slowdown is typically a longer wait time at the terminal. If the long wait time is not acceptable, the designer may have to increase the buffer pool size N. The tradeoff between cost of buffer storage and delay at the terminal must be weighed for a compromise solution.

Acknowledgment

The authors are grateful to C. Sauer for his valuable help in furnishing simulation results.

Appendix—distribution of the first message service time in slowdown period S_4, when processing time is constant.

For a stable system, it is required that $N(t_2) < N - L$. Assume that there are k arrivals during S_3. Two different cases may exist:

1. $k + N(t_2) \geq N - L$.

In this case, the normal period is terminated during S_3. Thus, $S_4 = S_3 - T_3$. Note that S_3 is a fixed value and T_3 is the $[N - L - N(t_2)]$th order sample of a uniform random variable over the interval $[0, S_3]$. Therefore, the distribution of S_4 can easily be derived.

2. $k + N(t_2) < N - L$.

Let the age of the last service time in T_3 be γ; hence $\gamma = d - S_4$. Since d is a constant, it suffices to obtain the distribution of γ. Suppose that $N(t_2 + S_3) = i < (N - L)$. The problem can be cast in the following fashion: What

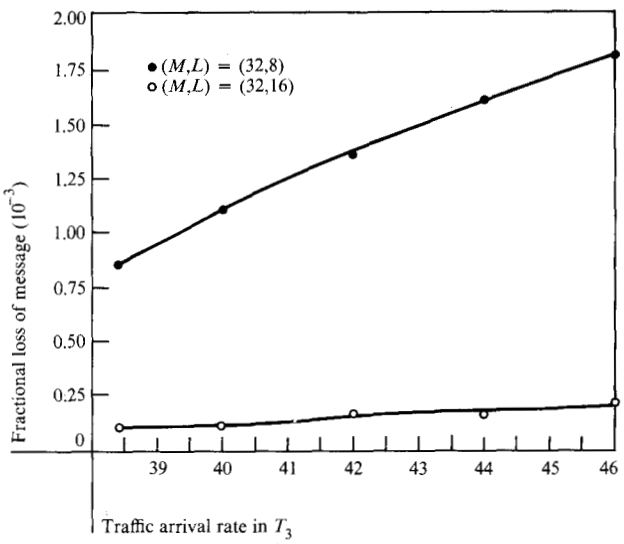

Figure 6 Fractional loss of messages vs traffic arrival rate in T_3.

is the age distribution of the last service time at which an M/D/1 queue starts its operation with queue length i and terminates its service at the first time the queue length is equal to $N - L$?

First, we investigate the conditional distribution of the queue length at the beginning of the last service, or $N(h^+)|h = t_3 - \gamma$, for a given i. (See Fig. 7.)

Define a Markov chain $\{X_n, n = 1, 2, \cdots\}$ such that X_n is the queue length in the system when the nth customer

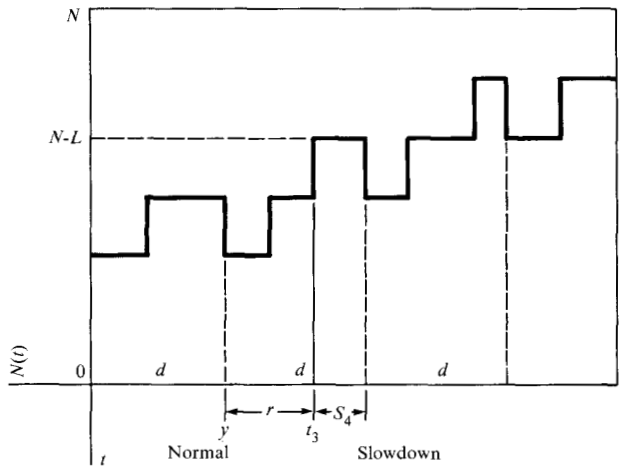

Figure 7 Number of used buffer units and age of the last service time in T_3.

is about to be served. $\{X_n\}$ can take one of the states $\{1, 2, \cdots, N-L-2, g\}$ where g is an absorbing state indicating $X_n \geq N-L-1$. Let

$$P_{ij}^{(m)} = P[X_{n+m} = j | X_n = i],$$

$$P_{ij} = P_{ij}^{(1)},$$

$y = t_3 - \gamma$, and

$$Q_j(i) = P[N(y^+) = j | N(t_2 + S_3) = i,\ N(t_3^+) = N-L].$$

Then

$$Q_j(i) = \sum_{n=1}^{\infty} P[X_{n-1} = j,\ X_n = g | X_1 = i].$$

Given $X_1 = i$, the event $\{X_{n-1} = j,\ X_n = g\}$ implies that $X_m \neq g$, $\forall m < n-1$, since the period T_3 would be terminated at any $X_m = g$. Using the Markov Property,

$$\begin{aligned} &P[X_{n-1} = j,\ X_n = g | X_1 = i] \\ &= P[X_n = g | X_{n-1} = j]\ P[X_{n-1} = j | X_1 = i] \\ &= P_{jg} P_{ij}^{(n-1)}. \end{aligned}$$

Therefore,

$$\begin{aligned} Q_j(i) &= \sum_{n=1}^{\infty} P_{jg} P_{ij}^{(n-1)} \\ &= \begin{cases} P_{jg} \sum_{n=1}^{\infty} P_{ij}^{(n)}, & i \neq j; \quad (31) \\ P_{ig}\left(1 + \sum_{n=1}^{\infty} P_{ij}^{(n)}\right), & i = j. \quad (32) \end{cases} \end{aligned}$$

Write

$$\begin{aligned} f_{ij} &= \sum_{n=1}^{\infty} P_{ij}^{(n)} \\ &= \sum_{k \neq j,g} P_{ik} f_{kj} + P_{ij}(1 + f_{ij}) \\ &= \sum_{k \neq g} P_{ik} f_{kj} + P_{ij},\ \forall i, j \neq g. \end{aligned} \quad (34)$$

Using Eqs. (31) – (34), $Q_j(i)$, $j = 1, 2, \cdots, N-L, 2$ can be solved.

To evaluate the age distribution of service γ, we consider first the conditional probability

$$\begin{aligned} &P[\gamma \leq t | N(h^+) = j,\ N(t_2 + S_3) = i] \\ &\qquad = F(t | j, N-L) / F(d | j, N-L), \end{aligned}$$

where

$$F(t | j, n) = \int_0^t e^{-\lambda y} \frac{(\lambda y)^{n-j-1}}{(n-j-1)!} \lambda dy.$$

Then, we have

$$P[\gamma \leq t | N(t_2 + S_3) = i] = \sum_{j=1}^{N-L-2} \frac{F(t | j, N-L)}{F(d | j, N-L)} Q_j(i). \quad (35)$$

After making Eq. (35) unconditional by using Eq. (24), the age distribution $P[\gamma \leq t]$ follows.

References

1. *Introduction to the IBM 3704 and 3705 Communication Controllers*, GA27-3051, IBM Corporation, White Plains, NY, 1974.
2. *IBM 3704 and 3705 Communication Controllers Network Control Program, Storage and Performance Estimates*, GC30-3006, IBM Corporation, White Plains, NY, 1974.
3. *IBM 3704 and 3705 Communication Controllers Network Control Program/VS Generation and Utilities*, GC30-3008, IBM Corporation, White Plains, NY, 1974.
4. *IBM 3704 and 3705 Communication Controllers, Network Control Program/VS Program Logic Manual*, SY30-3007-0, IBM Corporation, White Plains, NY, 1974.
5. S. M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, Inc., San Francisco, 1970.
6. N. U. Prabhu, *Queues and Inventories*, John Wiley & Sons, Inc., New York, 1965.
7. C. H. Sauer, "Simulation Analysis of General Queuing Network," *Summer Simulation Conference*, San Francisco, CA, 1975, p. 75.

Received September 22, 1976; revised December 7, 1976

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.