P. E. Boudreau
M. Kac

Analysis of a Basic Queuing Problem Arising in Computer Systems

Abstract: A model which describes a basic junction, or queuing structure, arising in a general computing system is subjected to a mathematical analysis. The results consist of several formulas describing the performance of various parts of the system. The feasibility in analyzing general queuing problems in this manner is stated, together with the results of a Monte Carlo simulation used for comparison purposes.

1. Introduction

Modern computers and processing systems have posed many new problems of system evaluation. The desirability of making the original source of information and the input source to the computer coincide, coupled with efforts in maximizing the use of all components of the system, has accented the queuing aspects of such problems. The standard methods of attacking the problems have been Monte Carlo simulations, and the use of elementary queuing theory, where exponential service times or Poisson distribution of arrivals are assumed. These methods have been used to attempt to answer such questions as:

1. What should be the relationship of input/output speeds to central processing speeds?
2. What is the size and type of buffer needed to handle queues that may develop in the system?
3. What action should be taken with regard to these queues?

The effort involved in simulations to answer these questions makes an analytic approach more attractive. However, the number of parameters that may affect the answers to such questions is so large that standard queuing techniques can, at best, provide only a general indication of what one may expect in a realistic situation.

This paper will study these problems by using a simplified model of the system rather than simplifying conditions external to the system. The sensitivity to various parameters may be studied on such a simplified model, and the feasibility of a complete analytic solution in the general case can be discussed.

The method of attack here is to have the simplified system represent a basic junction in any total system. At the very least, every processing system must perform the functions assumed in the model.

The problem under study is analogous to the problem of handling patients in a doctor's office. An input generator generates families of transactions in a given cycle time. (A secretary schedules appointments with the doctor.) The transactions are stored in a buffer (patients' waiting room). An output computer takes transactions from the buffer and processes them serially (the doctor treats patients one at a time). In the case of the simplified model, it is initially assumed that the output computer needs two cycles to process a single transaction (the doctor allows a fixed amount of time for each patient). Other assumptions which are made: first, the size of a given family of transactions is subject to a known probability distribution; second, the probability of the family having zero members is not zero; and third, the buffer is of unlimited capacity. The last assumption, though unrealistic, is not as restrictive as it may sound since one can still study fluctuations in the number of transactions waiting and the frequency with which overflows will occur.

The analysis is set up as a simple Markov chain. The results are that the utilization of the output computer and the average number of transactions waiting are both dependent on the average family size. If the process time is twice the generation time and the average generation size is less than 0.5, then the probability that the output computer is active at a given instant is twice the average generation size. If the average generation size is greater than 0.5, then the probability that the output computer is active at a given instant is 1. The average number of words in the buffer can be calculated using formulas of Section 4, Part C. The average time to the first overflow of a fixed capacity buffer can be calculated using formulas of Section 4, Part B. When the output computer needs k times the cycle time to process one transaction, the results are similar, though the underlying Markov chains become more complex and the analysis more tedious.

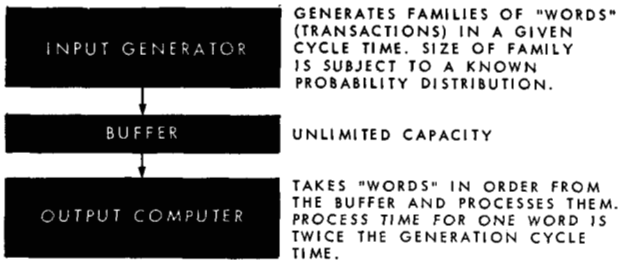

Figure 1 **Block diagram of mathematical model.**

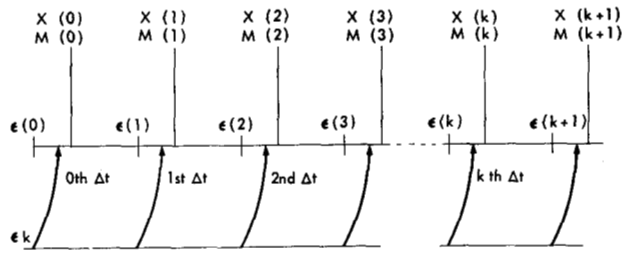

Figure 2 **Timing chart.**

Monte Carlo routines have been run on an IBM 704 to simulate the model, and the results compare quite favorably with those predicted by the theory.

2. The model

A block diagram of the model is shown in Figure 1. The input generator will produce in a cycle time a number of transactions or words. These words will be stored in the buffer. The output computer will take words from the buffer and process them one at a time. The time to process a single word will be two cycles.

• *Definitions*

Δt is the cycle time for the input generator.

k is the cycle number.

$\varepsilon(k)$ defines the status of the output computer at the k^{th} cycle time.

$$\varepsilon(k) = \begin{cases} 1 \text{ if output computer is busy at } k^{\text{th}} \text{ cycle time.} \\ 0 \text{ if output computer is not busy at } k^{\text{th}} \text{ cycle time.} \end{cases}$$

ρ_i is the probability that a given family of words has size i. It is assumed that $\rho_0 \neq 0$.
For consistency of notation $\rho_j = 0$ for $j < 0$.

$M(k)$ denotes the number of words in the buffer at the k^{th} cycle.

$X(k)$ denotes the size of the k^{th} family.

It is assumed that $X(0), X(1), \ldots$ are *independent random variables,* each subject to the distribution described by the ρ's.

• *Description of system status*

The process time for a word is assumed to be $(2-\varepsilon)\Delta t$ where ε can be arbitrarily small. This allows $\varepsilon\Delta t$ to compute $X(k)$, $M(k)$ and $\varepsilon(k)$ prior to the start of the k^{th} cycle.

The timing chart shown in Fig. 2 describes how the model functions. Because it takes slightly less than $2\Delta t$ to process a word, the value of $\varepsilon(k+1)$ depends entirely on $\varepsilon(k)$ and $M(k)$.[1]

$$\varepsilon(k+1) = \begin{cases} 1, \text{ if and only if } M(k) > 0 \text{ and } \varepsilon(k) = 0 \\ 0, \text{ if and only if } M(k) = \text{ or } \varepsilon(k) = 1 \end{cases} . \quad (1)$$

The number of words in the buffer at the $(k+1)^{\text{st}}$ Δt depends on the number of words in the buffer at the k^{th} Δt, the number of words generated at the $(k+1)^{\text{st}}$ Δt and whether or not a word was taken out of the buffer during the k^{th} cycle. This dependence is expressed by formula

$$M(k+1) = \text{Max}\{M(k) - [1-\varepsilon(k)], 0\} + X(k+1) . \quad (2)$$

The state of the system is completely defined at the k^{th} Δt by $M(k)$ and $\varepsilon(k)$.

Let $l = M(k)$

$\delta = \varepsilon(k)$

and (l, δ) denote the state of the system at the k^{th} cycle.

If (λ, n) is another state of the system, the transition probability from state (l, δ) to state (λ, n) in one cycle is denoted by $P(l, \delta | \lambda, n)$.

It is clear from Equation (2) and with our definition of the state that we are dealing with a Markov chain. It is here that the assumption of independence of the X's plays a critical part. Transition probabilities are:

$$P(l, \delta | \lambda, 1) = \begin{cases} \text{prob.}\{\text{Max}[l-(1-\delta), 0] + X(k) = \lambda\} \\ \qquad \text{if } l > 0 \text{ and } \delta = 0 \\ 0, \text{ otherwise} \end{cases}$$

and

$$P(l, \delta | \lambda, 0) = \begin{cases} \text{prob.}\{\text{Max}[l-(1-\delta), 0] + X(k) = \lambda\} \\ \qquad \text{if } l = 0, \text{ or } \delta = 1 \\ 0, \text{ otherwise .} \end{cases}$$

More specifically,

(a) $P(0, 0 | \lambda, 1) = 0$

(b) $P(0, 1 | \lambda, 1) = 0$

(c) $P(l, 0 | \lambda, 1) = \rho_{\lambda-(l-1)} \qquad l > 0$

(d) $P(l, 1 | \lambda, 1) = 0 \qquad l > 0$

(e) $P(0, 0 | \lambda, 0) = \rho_\lambda$

(f) $P(0, 1 | \lambda, 0) = \rho_\lambda$

(g) $P(l, 0 | \lambda, 0) = 0 \qquad l > 0$

(h) $P(l, 1 | \lambda, 0) = \rho_{\lambda-l} \qquad l > 0$

These are evident from the cycle definition chart shown in Fig. 3. For example if $l=0$, $\delta=0$ then n must be zero, which proves (a).

3. Recurrence relations and generating functions

• *A. Transition probability*

Of considerable interest is the transition probability from state (l, δ) to state (λ, n) in s steps.

Let $P^{(s+1)}(l, \delta|\lambda, n)$ = probability of going from state (l, δ) to state (λ, n) in $s+1$ steps. Since the process is Markovian, we have the recurrence relation

$$P^{(s+1)}(l, \delta|\lambda, n) = \sum_{\text{all } \lambda', n'} P^{(s)}(l, \delta|\lambda', n')P(\lambda', n'|\lambda, n) .$$

Renaming the variables gives:

$$P^{(s+1)}(l_0, \delta_0|\lambda, n) = \sum_{\text{all } l, \delta} P^{(s)}(l_0, \delta_0|l, \delta)P(l, \delta|\lambda, n) . \tag{3}$$

For the sake of brevity l_0 and δ_0 will be omitted from subsequent formulas so that, e.g., Formula (3) becomes

$$P^{(s+1)}(\lambda, n) = \sum_{l,\delta} P^{(s)}(l, \delta)P(l, \delta|\lambda, n) .$$

If $n=1$,

$$P^{(s+1)}(\lambda, 1) = \sum_{l,\delta} P^{(s)}(l, \delta)P(l, \delta|\lambda, 1)$$

and if $n=0$,

$$P^{(s+1)}(\lambda, 0) = \sum_{l,\delta} P^{(s)}(l, \delta)P(l, \delta|\lambda, 0) .$$

Using the one-cycle transition probabilities (a) through (h) we have:

$$P^{(s+1)}(\lambda, 1) = \sum_{l=1}^{\infty} P^{(s)}(l, 0)\rho_{\lambda-(l-1)} \tag{4}$$

and

$$P^{(s+1)}(\lambda, 0) = P^{(s)}(0, 0)\rho_\lambda + P^{(s)}(0, 1)\rho_\lambda + \sum_{l=1}^{\infty} P^{(s)}(l, 1)\rho_{\lambda-l} . \tag{5}$$

Next, we define the generating functions

$$F_1^{(s)}(z) = \sum_{\lambda=0}^{\infty} P^{(s)}(\lambda, 1)z^\lambda$$

$$F_0^{(s)}(z) = \sum_{\lambda=1}^{\infty} P^{(s)}(\lambda, 0)z^{\lambda-1} \tag{6}$$

$$f(z) = \rho_0 + z\rho_1 + z^2\rho_2 + \cdots = \sum_{j=0}^{\infty} \rho_j z^j . \tag{7}$$

Multiply both sides of Equation (4) by z^λ and sum on λ from 0 to ∞.

This gives:

$$\sum_{\lambda=0}^{\infty} P^{(s+1)}(\lambda, 1)z^\lambda = \sum_{l=1}^{\infty} P^{(s)}(l, 0) \sum_{\lambda=0}^{\infty} \rho_{\lambda-(l-1)}z^\lambda ,$$

Figure 3 **Cycle definition chart.**

which simplifies to

$$\sum_{\lambda=0}^{\infty} P^{(s+1)}(\lambda, 1)z^\lambda = f(z) \sum_{l=1}^{\infty} P^{(s)}(l, 0)z^{l-1} . \tag{8}$$

Referring to Equation (6) we have:

$$F_1^{(s+1)}(z) = f(z)F_0^{(s)}(z) . \tag{9}$$

Similarly from Equation (5) we obtain:

$$\sum_{\lambda=1}^{\infty} P^{(s+1)}(\lambda, 0)z^{\lambda-1} = P^{(s)}(0, 0) \sum_{\lambda=1}^{\infty} \rho_\lambda z^{\lambda-1} + P^{(s)}(0, 1) \sum_{\lambda=1}^{\infty} \rho_\lambda z^{\lambda-1} + \sum_{l=1}^{\infty} P^{(s)}(l, 1) \sum_{\lambda=1}^{\infty} \rho_{\lambda-l}z^{\lambda-1} ,$$

which simplifies to:

$$\sum_{\lambda=1}^{\infty} P^{(s+1)}(\lambda, 0)z^{\lambda-1} = P^{(s)}(0, 0)\frac{f(z)-\rho_0}{z} - P^{(s)}(0, 1)\frac{\rho_0}{z} + \frac{f(z)}{z} \sum_{\lambda=0}^{\infty} P^{(s)}(\lambda, 1)z^\lambda , \tag{10}$$

or

$$F_0^{(s+1)}(z) = P^{(s)}(0, 0)\frac{f(z)-\rho_0}{z} - P^{(s)}(0, 1)\frac{\rho_0}{z} + \frac{f(z)}{z} F_1^{(s)}(z) . \tag{11}$$

Now let,

$$G_1(z, w) = \sum_{s=0}^{\infty} F_1^{(s)}(z)w^s$$

$$G_0(z, w) = \sum_{s=0}^{\infty} F_0^{(s)}(z)w^s . \tag{12}$$

Multiplying both sides of Equation (9) by w^s and summing on s from 0 to ∞ we obtain

$$\frac{1}{w} \sum_{s=0}^{\infty} F_1^{(s+1)}(z)w^{s+1} = f(z) \sum_{s=0}^{\infty} F_0^{(s)}(z)w^s$$

or, by use of Equation (12):

$$\frac{1}{w}[G_1(z, w) - F_1^{(0)}(z)] = f(z)G_0(z, w) .$$

However,

$$F_1^{(0)}(z) = \sum_{\lambda=0}^{\infty} P^{(0)}(\lambda, 1)z^\lambda = 0$$

because $P^{(0)}(\lambda, 1)=0$ for all λ.

Thus we have:

$$G_1(z, w)=wf(z)G_0(z, w)\,. \tag{13}$$

Similarly from Equation (11)

$$\frac{1}{w}\sum_{s=0}^{\infty}F_0^{(s+1)}(z)w^{s+1}=\frac{f(z)-\rho_0}{z}\sum_{s=0}^{\infty}P^{(s)}(0, 0)w^s - \frac{\rho_0}{z}\sum_{s=0}^{\infty}P^{(s)}(0, 1)w^s + \frac{f(z)}{z}\sum_{s=0}^{\infty}F_1^{(s)}(z)w^s\,,$$

and again using Equation (12) we have:

$$\frac{1}{w}[G_0(z, w)-F_0^{(0)}(z)]=\frac{f(z)-\rho_0}{z}\sum_{s=0}^{\infty}P^{(s)}(0, 0)w^s - \frac{\rho_0}{z}\sum_{s=0}^{\infty}P^{(s)}(0, 1)w^s + \frac{f(z)}{z}G_1(z, w)\,.$$

However, now

$$F_0^{(0)}(z)=\sum_{\lambda=1}^{\infty}P^{(0)}(\lambda, 0)z^{\lambda-1}=\sum_{\lambda=1}^{\infty}\rho_\lambda z^{\lambda-1}=\frac{f(z)-\rho_0}{z}\,,$$

and we obtain

$$G_0(z, w)=\frac{f(z)-\rho_0}{z}+w\left[\frac{f(z)-\rho_0}{z}\sum_{s=0}^{\infty}P^{(s)}(0, 0)w^s - \frac{\rho_0}{z}\sum_{s=0}^{\infty}P^{(s)}(0, 1)w^s+\frac{f(z)}{z}G_1(z, w)\right]. \tag{14}$$

Setting

$$A(w)=\sum_{s=0}^{\infty}P^{(s)}(0, 0)w^s\,, \qquad B(w)=\sum_{s=0}^{\infty}P^{(s)}(0, 1)w^s\,, \tag{15}$$

and eliminating $G_1(z, w)$ from Eqs. (13) and (14), we obtain:

$$G_0(z, w)=\frac{\{w[A(w)]+1\}[f(z)-\rho_0]-w\rho_0B(w)}{z-w^2f^2(z)}\,. \tag{16}$$

We now wish to relate $A(w)$ to $B(w)$.

From Equation (5) we have:

$$P^{(s+1)}(0, 0)=P^{(s)}(0, 0)\rho_0+P^{(s)}(0, 1)\rho_0+\sum_{l=1}^{\infty}P^{(s)}(l, 1)\rho_{0-l} =P^{(s)}(0, 0)\rho_0+P^{(s)}(0, 1)\rho_0\,, \tag{17}$$

(since $\rho_j=0$ for $j<0$),

and taking generating functions we obtain from Eq. (17)

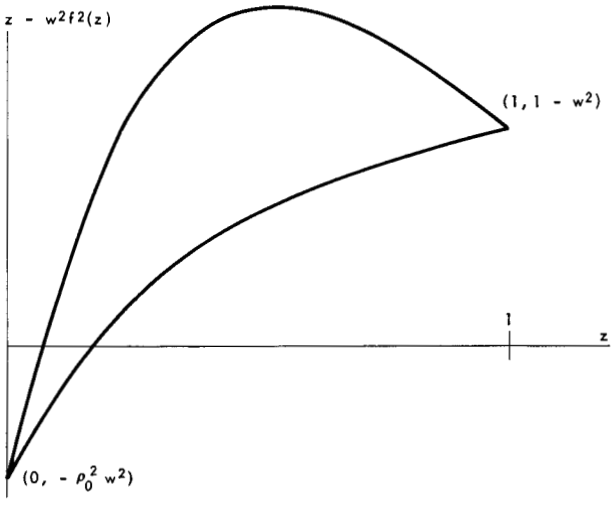

Figure 4 **Number of real roots between 0 and 1 for Eq. (21).**

$$\frac{1}{w}\sum_{s=0}^{\infty}P^{(s+1)}(0, 0)w^{s+1}=\rho_0\sum_{s=0}^{\infty}P^{(s)}(0, 0)w^s+\rho_0\sum_{s=0}^{\infty}P^{(s)}(0, 1)w^s\,,$$

or

$$\frac{1}{w}\left[\sum_{s=0}^{\infty}P^{(s)}(0, 0)w^s-P^{(0)}(0, 0)\right]=\rho_0A(w)+\rho_0B(w)\,.$$

Since $P^{(0)}(0, 0)=\rho_0$ we have

$$\frac{1}{w}[A(w)-\rho_0]=\rho_0[A(w)+B(w)]\,. \tag{18}$$

Finally,

$$G_1(z, w)=wf(z)G_0(z, w)\,, \tag{19}$$

and

$$G_0(z, w)=\frac{f(z)+wA(w)f(z)-A(w)}{z-w^2f^2(z)}\,. \tag{20}$$

B. Determination of A(w) from structure of G's

We have used all the information available from the recursions. However, $A(w)$ must still be determined. This may be done by analyzing the structure of $G_0(z, w)$. For every $w(0\leq w\leq 1)$, $G_0(z, w)$ is an analytic function of z in the open unit circle. Consider the equation:

$$z-w^2f^2(z)=0 \qquad 0<w<1\,. \tag{21}$$

How many real roots does Equation (21) have between 0 and 1 for given w?

Since $\frac{d}{dz}[z-w^2f^2(z)]=1-2w^2f(z)f'(z)$, this deriva-

tive is zero only if $f(z)f'(z)=\frac{1}{2w^2}$. For $0\leq z<1$, $f(z)f'(z)$ is monotonically increasing hence, at most, one critical point of $z-w^2f^2(z)$ can lie in (0, 1).

Because $\frac{d^2}{dz^2}[z-w^2f(z)]<0$ for $0\leq z<1$, $z-w^2f^2(z)$ has either a unique maximum, or no critical points between 0 and 1.

In either case, the graph in Fig. 4 shows there is a unique root of $z-w^2f^2(z)=0$ between 0 and 1. Let $\theta(w)$ denote this root.

Since $G_0(z, w)$ is analytic in the open unit circle, we must also have:

$$f[\theta(w)]+wA(w)f[\theta(w)]-A(w)=0\,,$$

or

$$A(w)=\frac{f(\theta)}{1-wf(\theta)}\,. \tag{22}$$

4. Formulas

• *A. Output computer utilization*

It is clear from the definition of $F_1^{(s)}(z)$ that, $F_1^{(s)}(1)$ is the probability that the output computer is being used at time s.

We shall calculate

$$\alpha=\lim_{n\to\infty}\frac{\sum_{s=0}^{n}F_1^{(s)}(1)}{n}=\lim_{n\to\infty}\frac{\sum_{s=0}^{n}\text{prob. }[\varepsilon(s)=1]}{n}$$

which represents the average probability that the output computer is in use.

Intuitively, this average probability also represents the percentage of time during which the output computer is used. A rigorous justification of this equivalence can be based on an appropriate "law of large numbers." However, we shall omit the justification and henceforth identify α with this percentage of time.[2]

The calculation of α, as well as the existence of the limit which defines it, is based on a tauberian theorem.

Accordingly we must find

$$\lim_{w\to1}(1-w)G_1(1, w)=\alpha\,.$$

We write

$$\lim_{w\to1}(1-w)G_1(1, w)=\lim_{w\to1}(1-w)wf(1)G_0(1, w)$$

$$=\lim_{w\to1}\left\{\frac{w(1-w)}{1-w^2}-\frac{w(1-w)f[\theta(w)]}{[1+w]\{1-wf[\theta(w)]\}}\right\}$$

$$=\lim_{w\to1}\frac{w}{w+1}\left\{1-\frac{(1-w)f[\theta(w)]}{1-wf[\theta(w)]}\right\}$$

and the question is: What is

$$\lim_{w\to1}\frac{(1-w)f(\theta)}{1-wf[\theta(w)]}\,?$$

From $\theta(w)-w^2f^2[\theta(w)]=0$, we obtain by differentiation

$$\theta'(w)=\frac{2wf^2[\theta(w)]}{1-2w^2f[\theta(w)]f'[\theta(w)]}\,. \tag{23}$$

We now consider three cases:

• *Case 1* $f'(1)<\frac{1}{2}$

In this case we shall prove that $\theta(w)\to1$ as $w\to1$. Suppose to the contrary, i.e., $\theta(w)\to\gamma<1$ as $w\to1$.

The graph of $z-w^2f^2(z)$ in this case is shown in Fig. 5.

It is clear that $\frac{d}{dz}[z-w^2f^2(z)]\neq0$, $0\leq z\leq1$

and there is no maximum.

Let β be such that $\gamma<\beta<1$.

Then $\beta-w^2f^2(\beta)\neq0$ for all w.

However, $\beta-w^2f^2(\beta)\leq1-w^2$ for all w.

Letting $w\to1$ we obtain $\beta-f^2(\beta)\leq0$.

However, since $\beta>\theta(w)$ for all w, we have $\beta-w^2f^2(\beta)\geq0$,

hence $\beta-w^2f^2(\beta)=0$, a contradiction.

Now,

$$\lim_{w\to1}\frac{(1-w)f[\theta(w)]}{1-wf[\theta(w)]}=\lim_{w\to1}\frac{1-w}{1-wf[\theta(w)]}\cdot\lim_{w\to1}f[\theta(w)]$$

and since

$$\theta(w)\to1 \text{ as } w\to1,\quad \lim_{w\to1}f[\theta(w)]=1\,,$$

thus,

$$\lim_{w\to1}\frac{1-w}{1-wf[\theta(w)]}=\lim_{w\to1}\frac{1}{1+w\dfrac{1-f[\theta(w)]}{1-w}}\,,$$

and

$$\frac{1-f[\theta(w)]}{1-\theta(w)}=\frac{f(1)-f[\theta(w)]}{1-\theta(w)}=f'(\zeta)$$

$$\text{for some } \theta(w)<\zeta<1\,.$$

We have

$$\lim_{w\to1}\frac{1}{1+w\dfrac{1-f[\theta(w)]}{1-w}}=\lim_{w\to1}\frac{1}{1+w\dfrac{1-\theta(w)}{1-w}f'(\zeta)}$$

and clearly

$$\lim_{w\to1}f'(\zeta)=f'(1)\,.$$

Finally,

$$\lim_{w\to1}\frac{1-\theta(w)}{1-w}=\lim_{w\to1}\theta'(w)=\frac{2}{1-2f'(1)}$$

where we have made use of formula (23).

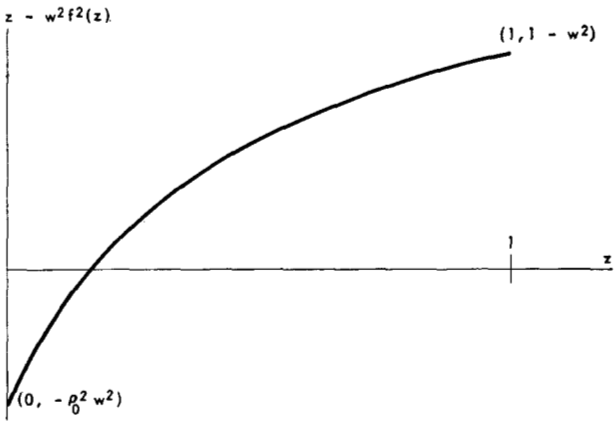

Figure 5 **Relationship between z and $z-w^2f^2(z)$.**

Thus,

$$\lim_{w\to 1}(1-w)G_1(1,w)=f'(1) \qquad (24)$$

and $\alpha=f'(1)$.

• *Case 2* $f'(1)>\frac{1}{2}$

First we show that $\theta(w)\to\gamma<1$ as $w\to 1$.

Since $f'(1)>\frac{1}{2}$, there is a maximum for $z-w^2f^2(z)$ in the interval $(0, 1)$ for each w. Thus there is a z_0 such that

$$f(z_0)f'(z_0)=\frac{1}{2w^2} \text{ for } w \text{ sufficiently close to 1.}$$

We know that $\theta(w)<z_0$ since $\theta(w)$ is unique and the maximum occurs at z_0.

For a given $f'(1)>\frac{1}{2}$, $\quad z_0<1$,

thus $\theta(w)\to\gamma$ where $\gamma<z_0<1$.

We conclude in this case that

$$\lim_{w\to 1}\left[\frac{w}{1+w}\left(1-\frac{(1-w)f[\theta(w)]}{1-wf[\theta(w)]}\right)\right]=\tfrac{1}{2},$$

hence

$$\lim_{w\to 1}(1-w)G_1(1,w)=\tfrac{1}{2}, \qquad (25)$$

thus $\alpha=\frac{1}{2}$.

• *Case 3* $f'(1)=\frac{1}{2}$

In this case a more elaborate analysis is needed; it is omitted because the case is too special to justify the amount of effort.

• *B. Average time to first overflow of a finite capacity buffer*

Let N be the capacity of the buffer, and $M(k)$ is as before the number of words in the buffer after the k^{th} generation.

Define

$$Q_{n+1}(g,\delta)=\text{Probability}\{M(0)<N, M(1)<N, \cdots M(n)<N, M(N+1)=g \text{ and } \varepsilon(n+1)=\delta\}.$$

That is, $Q_{n+1}(g,\delta)$ is the probability that the capacity of the buffer has not been exceeded through cycle n, and that there are exactly g words in the buffer at the $(n+1)^{\text{st}}$ Δt; further, the status of the output computer at the $(n+1)^{\text{st}}$ Δt is δ.

Thus

$$Q_{n+1}(g,\delta)=\sum_{\substack{m<N\\ \varepsilon}}\text{prob}\{M(0)<N, \cdots M(n)=m, \varepsilon(n)=\varepsilon\}\cdot$$

$$\text{prob}\{M(0)<N, \cdots, M(n)=m, \varepsilon(n)=\varepsilon|\cdot$$

$$M(n+1)=g \text{ and } \varepsilon(n+1)=\delta\},$$

or

$$Q_{n+1}(g,\delta)=\sum_{\substack{m<N\\ \varepsilon}}\text{prob}\{M(0)<N, \cdots, M(n-1)<N,$$

$$M(n)=m, \varepsilon(n)=\varepsilon\}P(m,\varepsilon|g,\delta).$$

Finally, we have the recursion

$$Q_{n+1}(g,\delta)=\sum_{\substack{m<N\\ \varepsilon}}Q_n(m,\varepsilon)\cdot P(m,\varepsilon|g,\delta). \qquad (26)$$

We are only interested in $g<N$, thus if we restrict m, g to the values

$m=0, 1, 2, \cdots, N-1 \qquad \varepsilon=0, 1$

$g=0, 1, 2, \cdots N-1 \qquad \delta=0, 1$

then we can write Equation (26) in matrix notation. Consider the $2n\times 2n$ matrix

$P\equiv P(\underbrace{m,\varepsilon}_{\text{row}}|\underbrace{g,\delta}_{\text{column}})$

$$\begin{array}{c} \\ (0,0)\\(1,0)\\ \vdots \\ (N-1,0)\\(0,1)\\ \vdots \\ (N-1,1)\end{array}\begin{array}{c}\begin{array}{cccccccc}(0,0) & (1,0) & \cdots & (N-1,0) & (0,1) & \cdots & (N-1,1)\end{array}\\ \left[\begin{array}{cccccccc} & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & P(m,\varepsilon|g,\delta) & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \end{array}\right]\end{array} \qquad (27)$$

Note that

$$Q_0(m,\varepsilon)=\text{Prob}\{m(0)=m, \varepsilon(0)=\varepsilon\}=P(0,0|m,\varepsilon),$$

and $\quad P(0,0|m,\varepsilon)=\begin{cases}\phi \text{ if } \varepsilon=1\\ \rho_m \text{ if } \varepsilon=\phi\end{cases}.$

$$Q_1(g,\delta)=\sum_{\substack{m<N\\ \varepsilon}}Q_0(m,\varepsilon)\cdot P(m,\varepsilon|g,\delta).$$

Write $Q_0(m, \varepsilon)$ for $m=0, 1, \cdots, N-1$, $\varepsilon=0, 1$ as a row vector:

$$Q_0=\{Q_0(0,0)Q_0(1,0), \cdots Q_0(N-1,0),$$
$$Q_0(0,1)Q_0(1,1) \cdots Q_0(N-1,1)\},$$

and likewise

$$Q_n=\{Q_n(0,0), Q_n(1,0), \cdots Q_n(N-1,0)Q_n(0,1) \cdots$$
$$Q_n(N-1,1)\}.$$

The recursion equation (26) can now be written in the form $Q_{n+1}=Q_nP$,

hence

$$Q_n=Q_0P^n. \tag{28}$$

Now

$$Q_n\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}=\text{Prob}\{M(0)<N, \cdots M(n)<N\}=P_n(N). \tag{29}$$

We have

$$P_n(N)=Q_n\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}=Q_0(P^n)\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}=Q_0P^n(1). \tag{30}$$

$$P_n(N)-P_{n+1}(N)=\text{Prob}\{M(0)<N, \cdots, M(n)<N,$$
$$M(n+1)\geqslant N\},$$

and equals the probability that the capacity N is exceeded for the first time at $n+1$.

By definition

Average time to first overflow

$$=\sum_{n=0}^{\infty}(n+1)[P_n(N)-P_{n+1}(N)]$$
$$=(P_0(N)-P_1(N))+2[P_1(N)-P_2(N)]$$
$$+3[P_2(N)-P_3(N)]+\cdots,$$

and if $nP_n(N)\to 0$,

Average time to first overflow

$$=P_0(N)+P_1(N)+P_2(N)+\cdots$$
$$=[Q_0(P)^0+Q(P)^1+Q_0(P)^2+\cdots](1)$$
$$=Q_0[I+(P)^1+(P)^2+\cdots](1)$$
$$=Q_0[I-(P)]^{-1}(1), \tag{31}$$

where I is the identity matrix.

Thus we have a computational scheme for determining the average time before first overflow. Similar techniques can be employed to answer more complex questions.

• *C. Average number of words in the buffer*

By the definitions of this family in Eq. (6) of $F_0^{(s)}(z)$ and $F_1^{(s)}(z)$ we know that

$$\frac{d}{dz}[F_1^{(s)}(z)+zF_0^{(s)}(z)]_{z=1}=\sum_{l=1}^{\infty}l[P^{(s)}(l,0)+P^{(s)}(l,1)].$$

However, $P^{(s)}(l,0)+P^{(s)}(l,1)$ is the probability that there are exactly l words in the buffer at time $s\Delta t$. Let $N(s)$ = average number of words in the buffer at time $s\Delta t$. Then

$$N_s=\frac{d}{dz}[F_1^{(s)}(z)+zF_0^{(s)}(z)]_{z=1}$$
$$=\sum_{l=1}^{\infty}l[P^{(s)}(l,0)+P^{(s)}(l,1)]. \tag{32}$$

Take a generating function of w on Eq. (32)

$$\sum_{s=0}^{\infty}N(s)w^s=\sum_{s=0}^{\infty}w^s\frac{d}{dz}[F_1^{(s)}(z)+zF_0^{(s)}(z)]_{z=1}$$
$$=\frac{d}{dz}[G_1(z,w)+zG_0(z,w)]_{z=1}$$
$$=\frac{d}{dz}\left\{[w+f(z)+z]\left[\frac{f(z)+\dfrac{wf(z)+f[\theta(w)]}{1-wf[\theta(w)]}-\dfrac{f[\theta(w)]}{1-wf[\theta(w)]}}{z-w^2f^2(z)}\right]_{z=1}\right\}$$
$$=\frac{d}{dz}[w+f(z)+z]\left[\frac{f(z)-f[\theta(w)]}{\{1-wf[\theta(w)]\}[z-w^2f^2(z)]}\right]_{z=1},$$

or,

$$\sum_{s=0}^{\infty}N(s)w^s=\frac{1-f[\theta(w)]}{1-w^2}-\frac{1-wf[\theta(w)]}{[1+wf'(1)]}$$
$$+\frac{1+w}{1-wf[\theta(w)]}\left\{\frac{f'(1)}{1-w^2}-\frac{\{1-f[\theta(w)]\}[1-zw^2f'(1)]}{(1-w^2)^2}\right\}. \tag{33}$$

If $f'(1)<\frac{1}{2}$ we find by an elementary but laborious calculation that

$$\sum_{s=0}^{\infty}N(s)w^s\approx\frac{f''(1)+f'(1)[1-f'(1)]}{(1-w)[1-zf'(1)]}. \tag{34}$$

Applying again the tauberian theorem we obtain — *Theorem:* If the average generation size, $f'(1)<\frac{1}{2}$

then

$$\frac{\sum_{s=0}^{n}N(s)}{n}\to\frac{f''(1)+f'(1)[1-f'(1)]}{1-2f'(1)}. \tag{35}$$

If $f'(1)>\frac{1}{2}$ one obtains

$$\frac{\sum_{s=0}^{n} N(s)}{n^2} \to \gamma , \tag{36}$$

where γ can be easily calculated.

In this case we see that $N(s)$ will have to become very large on occasions and frequent overflows will occur. However, we do not enter into a detailed discussion of this case since in most practical applications $f'(1)<\frac{1}{2}$, or can be made so.

Note that Equation (35) gives only the "time average" of the average number of words in the buffer. For better understanding, fluctuations of the number of words in the buffer should be studied in detail. This is entirely feasible but rather tedious.

5. Analysis for a process time three times greater than the generation cycle

Consider the case where the process time is $3\Delta t$.

The critical difference between this case and the $2\Delta t$ case is that the status of the output computer at time $k\Delta t$ depends on the status at $(k-2)\Delta t$ as well as $(k-1)\Delta t$.

Here,

$$\varepsilon(k+1)=\begin{cases} 1, \text{ if and only if} \\ M(k)>0 \text{ and } \varepsilon(k)=0 \\ \text{or} \\ M(k-1)>0 \text{ and } \varepsilon(k-1)=0 \\ 0, \text{ if and only if} \\ M(k)=0 \text{ and } M(k-1)=0 \\ \text{or} \\ M(k)=0 \text{ and } \varepsilon(k-1)=1 \\ \text{or} \\ \varepsilon(k)=1 \text{ and } M(k-1)=0 \\ \text{or} \\ \varepsilon(k)=1 \text{ and } \varepsilon(k-1)=1 \end{cases}$$

Four variables are needed to define the state of the model

$M(k)=l \qquad \varepsilon(k)=\delta$

$M(k-1)=l_1 \qquad \varepsilon(k-1)=\delta_1 .$

There are 16 single-cycle transition probabilities of the form

$$P(l, l_1, \delta, \delta_1, | \lambda, \lambda_1, n, n_1)$$

where

$\lambda=m(k+1)$

$\lambda_1=m(k)=l$

$n=\varepsilon(k+1)$

$n_1=\varepsilon(k)=\delta .$

An analysis similar to, but more complicated than, the one used to obtain Eqs. (24) and (25) yields

$$\lim_{n\to\infty} \frac{\sum_{s=0}^{n} \text{prob}\{\varepsilon(s)=1\}}{n} = 2f'(1) \qquad \text{if } f'(1)<\tfrac{1}{3} \tag{37}$$

and

$$\lim_{n\to\infty} \frac{\sum_{s=0}^{n} \text{prob}\{\varepsilon(s)=1\}}{n} = \tfrac{2}{3} \qquad \text{if } f'(1)>\tfrac{1}{3} . \tag{38}$$

The form of Eqs. (37) and (38) is so analogous to that of Eqs. (24) and (25) that one can surmise the following general theorem.

Theorem:

If the "speed" of the output computer is $k\Delta t$ then

$$\lim_{n\to\infty} \frac{\sum_{s=0}^{n} \text{prob}\{\varepsilon(s)=1\}}{n} = (k-1)f'(1) \qquad \text{if } kf'(1)<1 \tag{39}$$

and

$$\lim_{n\to\infty} \frac{\sum_{s=0}^{n} \text{prob}\{\varepsilon(s)=1\}}{n} = \frac{(k-1)}{k} \qquad \text{if } kf'(1)>1 . \tag{40}$$

Expressions analogous to G_0 and G_1 of Sec. III can be found and hence one can discuss fully the fluctuations of the number of words stored in the buffer.

6. Results of Monte Carlo methods

The results of the Monte Carlo routines run on the IBM 704 can be easily compared to the analytical results of Eqs. (24) and (35).

Five distributions on the ρ's have been considered, and ten runs for each distribution have been made, each run consisting of one-thousand cycles.

The following tables and Fig. 6 indicate the results of the simulation versus the analytical results.

Table 1 **The five cases.**

	ρ_0	ρ_1	ρ_2	ρ_3	$f'(1)$	$f''(1)$
Case 1	0.7	0.2	0.1	0.0	0.4	0.2
Case 2	0.95	0.025	0.0125	0.0125	0.0875	0.1
Case 3	0.85	0.05	0.05	0.05	0.3	0.4
Case 4	0.81	0.18	0.01	0.0	0.2	0.02
Case 5	0.7	0.2	0.05	0.05	0.45	0.4

Table 2 **Average number of words in buffer.**

Average of N(s)	*Case 1*	*Case 2*	*Case 3*	*Case 4*	*Case 5*
From Simulation	2.45	0.216	1.47	0.298	5.73
From Theory	2.2	0.218	1.525	0.3	6.475

The results indicate that as $f'(1)\rightarrow\frac{1}{2}$, the number of words in the buffer exhibits larger fluctuations. This is also predicted by theory since $1-2f'(1)\rightarrow 0$ as $f'(1)\rightarrow\frac{1}{2}$ (Eq. 35).

Table 3 **M(k) for k=0−99. Case 1.**

k	$M(k)$	k	$M(k)$	k	$M(k)$	k	$M(k)$
0	0	26	4	51	2	76	2
1	0	27	4	52	2	77	3
2	1	28	4	53	1	78	2
3	0	29	3	54	3	79	2
4	1	30	5	55	3	80	1
5	1	31	5	56	3	81	3
6	1	32	6	57	2	82	2
7	0	33	6	58	2	83	3
8	0	34	6	59	1	84	2
9	1	35	5	60	1	85	3
10	0	36	5	61	0	86	3
11	0	37	4	62	1	87	3
12	0	38	5	63	0	88	2
13	0	39	4	64	0	89	3
14	0	40	4	65	0	90	2
15	1	41	3	66	0	91	2
16	0	42	3	67	1	92	1
17	0	43	3	68	0	93	1
18	2	44	3	69	2	94	1
19	2	45	3	70	2	95	1
20	2	46	3	71	2	96	0
21	3	47	2	72	1	97	1
22	4	48	2	73	1	98	1
23	3	49	2	74	0	99	1
24	4	50	3	75	2		
25	4						

7. Conclusion

An analytic approach to problems posed in the Introduction is feasible and fruitful. Two important simplifications are introduced:

(a) The buffer is assumed to be of infinite capacity.
(b) The processing time is assumed to be a fixed multiple of the generation time.

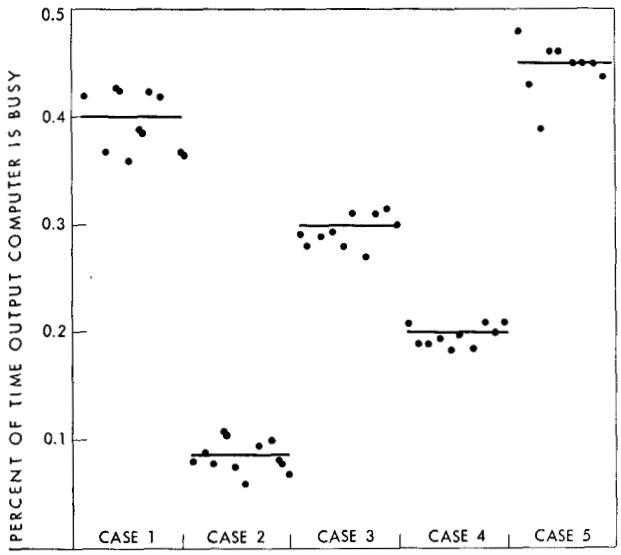

Figure 6 **Percent of time computer is busy.** *Horizontal line indicates the analytical $f'(1)$ while dots indicate results of simulation.*

The case when the processing time is a random multiple of the generation time (subject to a prescribed distribution function) is now under study, and we hope to present the results at a later date.

Should the capacity of the buffer be limited, an analytic approach is probably hopeless because there is no longer a tractable Markov chain.

Even though the assumption of infinite capacity is unrealistic, the study of the fluctuation of the number of words in the buffer should yield a better understanding of the actual processes as they occur in practice.

In particular, good estimates of the frequency with which overflows occur can certainly be obtained.

The satisfactory agreement between the theory presented and the results of Monte Carlo calculations should increase the belief in the accuracy of simulation calculations when applied to situations too complex to be amenable to analytic treatment.

Finally, it is hoped that this work will stimulate further thinking on the important problem of computer system evaluation.

Footnotes

1. Note the $\varepsilon(k)$ refers to the status of the model just before the k^{th} cycle begins; this is why the ε's on the chart are placed a little to the left of the vertical lines.
2. Actually it is 2α which is the percentage of time the output computer is used because for every "busy" cycle the following cycle is "not available."

Received June 28, 1960