The Distribution of Waiting Times in Clocked Multistage Interconnection Networks

CLYDE P. KRUSKAL, MARC SNIR, AND ALAN WEISS

Abstract—We analyze the random delay experienced by a message traversing a buffered, multistage packet-switching banyan network. We find the generating function for the distribution of waiting time at the first stage of the network for a very general class of traffic, assuming messages have discrete sizes. For example, traffic can be uniform or nonuniform, messages can have different sizes, and messages can arrive in batches. For light-to-moderate loads, we conjecture that delays experienced at the various stages of the network are nearly the same and are nearly independent. This allows us to approximate the total delay distribution. Better approximations for the distribution of waiting times at later stages of the network are attained by assuming that in the limit a sort of spatial steady state is achieved. Extensive simulations confirm the formulas and conjectures.

Index Terms—Banyan networks, discrete networks, interconnection networks, parallel processing, performance analysis, queueing theory.

I. INTRODUCTION

BUFFERED interconnection networks are receiving increasing consideration for use in parallel computers. They are integral components of several machines currently under development, including the Cedar machine at the University of Illinois, the NYU Ultracomputer at New York University [8], and the RP3 machine [16] at IBM, where they are used to interconnect processors to shared memory. In order to study the multitude of options available in actually building a machine, it is extremely useful to have formulas that approximate the performance of an interconnection network. In fact, formulas derived in a previous paper by two of the authors [11] have been heavily used in designing both the NYU Ultracomputer [8] and RP3 [14]. While simulation results are often more accurate, they are time consuming and expensive. In this paper, we analyze the random delay experienced by a message traversing a buffered, multistage packet-switching banyan network, for a very general class of traffic. For example, traffic can be uniform or nonuniform, messages can have different sizes, and messages can arrive in batches.

Interconnection networks connect processing elements to memory modules through stages of switches (Fig. 1). Early work in describing these networks was done by Goke and Lipovski [7], Lawrie [13], and Patel [15], among others. For fuller explanations of interconnection networks see [6] or [17], for example. There have been a number of performance analyses of interconnection networks (e.g., [4], [5], [11], [12]).

The basic building block of an interconnection network is a k-input, s-output ($k \times s$) buffered switch (Fig. 2). Each input port can accept one packet per clock cycle, and route it to the appropriate output port. Each output port has a FIFO buffer. Conflicts between messages simultaneously routed to the same output port are resolved by queueing the messages. We idealize this structure by assuming that the output buffers have infinite length. While this is clearly infeasible in practice, it is well known that for light-to-moderate loads, moderate-sized buffers provide approximately the same performance as infinite buffers. We also assume that each output port buffer contains any number of messages from the input ports in a clock cycle, and that arriving messages do not interfere with departing messages. Each output port can be viewed as a discrete queueing system.

We make the following probabilistic assumptions concerning traffic at the first stage of a network.

1) The number of messages arriving at successive cycles to an output port are independent, identically distributed random variables. These random variables may have a different distribution at different ports, and are clearly dependent from port to port.

2) The service requirements (the number of cycles required to forward a packet) for successive messages at an output port are independent, identically distributed random variables. This distribution may vary from port to port.

Constant service time is usually the appropriate assumption for interconnection networks realized with synchronous logic.

Assuming that the traffic is uniform (e.g., each request is equally likely to go to every output node) and that at each cycle a packet arrives at an input node with a fixed probability $p$, the expected delay has been computed [11]. This analysis is based on Little's identity, and it is not obvious that it can be extended to obtain more information about the delay distribution, such as the variance. Such information is quite important for two reasons. First, to obtain good performance on a parallel machine, it is not sufficient to have a low expected memory access time; high variance will impede performance, as it is often the case that the speed of the slowest processor dictates the system speed. Second, as we discuss in Section V, the
better approximations for the waiting time at the later stages. In Section V, we discuss how to analyze the total delay through the network. Section VI gives some concluding remarks.

II. Analysis

Our model for the first stage of switching comes under the general rubric of a discrete time queueing system. We compute in this section the $z$ transform for the waiting time, and use it to derive the expectation and variance of the waiting time, for general discrete service and arrival distributions. The solution method we used was indicated by Kobayashi and Konheim [lo]. We are not, however, aware of a complete solution to this problem in the literature. As will be seen in the remainder of this paper, for the queueing systems we are interested in, it is useful to carry out the calculations in their entirety.

We start with some definitions. Let

$$\lambda = \text{average number of arrivals at any cycle}$$

and

$$m = \text{average service time of a message.}$$

The traffic intensity is then

$$\rho = m\lambda.$$ 

Let

$$f_j = \text{probability that } j \text{ messages arrive at any cycle}$$

and

$$R(z) = \sum_{j=0}^{\infty} f_j z^j.$$ 

Then

$$R'(1) = \lambda.$$ 

Let

$$g_j = \text{probability that a message requires } j \text{ time units to serve}$$

and

$$U(z) = \sum_{j=0}^{\infty} g_j z^j.$$ 

Then

$$U'(1) = m.$$ 

Theorem 1: Let $w$ be the steady-state waiting time for a message. The $z$ transform of the waiting time distribution of an output queue is

$$t(z) = E(z^w) = \Psi(z) \phi(U(z))$$

$$= \frac{1 - m \lambda}{\lambda} \frac{1 - z}{1 - R(U(z))} \frac{1 - R(U(z))}{R(U(z)) - z} 1 - U(z).$$ (1)
Proof: Let \( s_n \) be the unfinished work at the end of the \( n \)th cycle, \( a_n \) be the number of messages arriving at the \( n \)th cycle, and \( c_n \) be the total service time for messages arriving at the \( n \)th cycle. Let \( s, a, \) and \( c \) be the steady-state variables corresponding to \( s_n, a_n, \) and \( c_n. \) Note that

\[
E(z^s) = E(z^a) = R(z),
\]

\[
E(z^c) = \sum_{j=0}^{\infty} E(z^c|a=j) f_j = \sum_{j=0}^{\infty} (U(z))^j f_j = R(U(z)),
\]

and

\[
s_n = \max(0, s_{n-1} + c_n - 1).
\]

Since \( c_n \) is independent of \( s_{n-1}, \) we obtain the identity

\[
E(z^c) = E(z^{c-s} | s > 0) P(s > 0)
\]

\[
+ E(z^{c-s} | s = 0, c > 0) P(s = 0, c > 0)
\]

\[
+ E(z^c | s = 0, c = 0) P(s = 0, c = 0)
\]

\[
= E(z^c) E(z^{c-s} | s > 0) P(s > 0)
\]

\[
+ E(z^{c-s} | c > 0) P(s = 0, c > 0)
\]

\[
+ P(s = 0) P(c = 0).
\]

Let

\[
\Psi(z) = \sum_{j=0}^{\infty} h_j z^j = E(z^c).
\]

The previous identity implies

\[
\Psi(z) = R(U(z)) \left( \frac{(z-1) R(U(0))}{z} + h_0 \frac{R(U(z)) - R(U(0))}{z} \right) + h_0 R(U(0))
\]

so that

\[
\Psi(z) = \frac{h_0 (1-z) R(U(0))}{R(U(z)) - z}.
\]

We compute, using L'Hospital's rule

\[
\Psi(1) = 1 = \frac{h_0 R(U(0))}{1 - m \lambda},
\]

so that

\[
t'(z) = E'(z^w) = \frac{1 - m \lambda}{\lambda} - \frac{U'(z) U(z) [1 - U(z)] [1 - z] + [1 - U(z)] [1 - R(U(z))] + [R(U(z)) - z] [1 - z] [1 - R(U(z))] U'(z)}{[R(U(z)) - z]^2 [1 - U(z)]^2}.
\]

Taking the limit as \( z \) approaches 1 (by applying L'Hospital's rule four times), we get

\[
t'(1) = E(w) = \frac{m R(z) U'(z)}{2 \lambda (1 - m \lambda)}.
\]
looking at the second derivative of $t(z)$:

$$
t''(1) = E(w(w-1)) = E(z^2)_{z=1} = \frac{(2R''(1)m^2 + 2\lambda^2 U''(1))(1 - m\lambda) + 3R''(1) m^2 + 3\lambda^2 U''(1)^2 + 3R''(1) U''(1) + 3R''(1) U''(1) m\lambda}{6\lambda(1 - m\lambda)^2}
$$

which yields

$$
Var(w) = E(w(w-1)) + E(w)(1 - E(w)) - \frac{(6m\lambda R''(1) + 4m^2 \lambda R''(1) + 6\lambda^2 U''(1) + 4\lambda^2 U''(1))(1 - m\lambda) - 3m^2 R''(1)(1 - 2m\lambda) + 3\lambda^4 U''(1) + 6R''(1) U''(1)}{12\lambda^2(1 - m\lambda)^2}.
$$

(The derivation of $t''(1)$ used six applications of L'Hospital's rule, and took Macsyma all night on a minicomputer.)

### III. EXAMPLES

We now apply the above formulas to derive the expected value and the variance of the waiting time for messages in several standard and important queueing systems. Note that the expected value formulas only give the waiting time of a message. To obtain the delay of a message in a queue, one must add to these formulas the service time. For the queueing systems in this section, message arrivals are independent of queue length. Thus, the variance of the delay of a message in a queue is simply the sum of the variance of the waiting time and the variance of the service time.

#### A. Service Time One

Suppose that all messages take exactly one unit of time to be serviced. Then $m = 1$ and $U(z) = z$. Thus,

$$
U'(z) = 1
$$

and

$$
U''(z) = U'''(z) = 0.
$$

Substituting into (1), we obtain

$$
t(z) = E(z^2) = \frac{1 - \lambda - R(z)}{\lambda R(z) - z}.
$$

Substituting into (2), we get

$$
E(w) = \frac{R'(1)}{2\lambda(1 - \lambda)}
$$

and substituting into (3), we get

$$
Var(w) = \frac{2(3R''(1) + 2R'''(1))\lambda(1 - \lambda) - 3(1 - 2\lambda)(R''(1))^2}{12\lambda^2(1 - \lambda)^2}.
$$

We analyze some special cases of this for $k$-input, $s$-output $(k \times s)$ switches.

1) Uniform Traffic, Single Arrivals: Suppose that each input port has a probability $p$ of receiving one message at each unit of time, and that each incoming message has an equal chance of going to any of the output ports. Then

$$
f_i = \binom{k}{j} \left( \frac{p}{s} \right)^j \left( 1 - \frac{p}{s} \right)^{k-j}.
$$

This quickly yields

$$
R(z) = \left( \frac{1 - \frac{p}{s} + \frac{pz}{s} \lambda}{s} \right)^k
$$

so that

$$
R'(1) = \lambda = \frac{kp}{s},
$$

and

$$
R''(1) = k(k-1) \left( \frac{p}{s} \right)^2 = \lambda^2 \left( 1 - \frac{1}{k} \right),
$$

Hence, substituting into (4) and (5),

$$
E(w) = \frac{\left( 1 - \frac{1}{k} \right) \lambda}{2(1 - \lambda)}
$$

and

$$
Var(w) = \frac{\left( 1 - \frac{1}{k} \right) \lambda \left[ 6 - 5\lambda \left( 1 + \frac{1}{k} \right) + 2\lambda^2 \left( 1 + \frac{1}{k} \right) \right]}{12(1 - \lambda)^2}.
$$

2) Bulk Arrivals: In many systems, the size of a message exceeds the size of a transmission packet; a message is transmitted in several packets. These packets arrive at the first stage of the network in one bulk. This can be modeled as in the previous example, except arrivals at input ports are in batches.
Assuming a constant batch size of \( b \) messages,
\[
R(z) = \left( 1 - \frac{p}{s} + \frac{pz^b}{s} \right)^k .
\]

Thus,
\[
R'(1) = \frac{kbp}{s} ,
\]
and
\[
R^*(1) = \frac{b kp}{s} (b-1 + \frac{b(k-1)p}{s}) .
\]

Using (4) and (5), this gives
\[
E(w) = \frac{(b-1) + \left( 1 - \frac{1}{k} \right) \lambda}{2(1 - \lambda)}
\]
and
\[
\text{Var}(w) = \frac{b^2 - 1 + 2\lambda \left( b^2 + 2 - \frac{3b}{k} \right) - 5\lambda^2 \left( 1 - \frac{1}{k^2} \right) + 2\lambda^3 \left( 1 - \frac{1}{k^2} \right)}{12(1 - \lambda)^2} .
\]

These agree with our previous formulas for the case \( b = 1 \).

3) Nonuniform Traffic: In many practical situations, each input is likely to have a distinct favorite output port (e.g., the output port connecting a processor to its private memory—see [1]). We assume that \( k = s \). (It is not hard to generalize this for \( k \neq s \), but the equations become quite lengthy.) We do assume bulk arrivals. Each input port sends arriving messages to its favorite output port with probability \( q \), and sends them with probability \((1 - q)/k\) to each output port (including its favorite output). The distribution of messages at the output ports is the product of two terms: the first term accounts for normal messages and is essentially the same as given in Section III-A-2, with \( p \) replaced by \( p(1 - q) \); the second term accounts for favored messages. We get
\[
R(z) = \left( 1 - p \frac{1-q}{k} + p \frac{1-q}{k} z^p \right)^{k-1} \cdot \left( 1 - p \left( q + \frac{1-q}{k} \right) + p \left( q + \frac{1-q}{k} \right) z^b \right) .
\]

Thus,
\[
R'(1) = \lambda = bp ,
\]
and
\[
R^*(1) = bp \left( \frac{1}{k} \left( 1 - q^2 \right) + (b-1) \right) .
\]

Note that for \( q = 1 \), we get \( E(w) = 0 \), and for \( q = 0 \) we obtain the same formula as in Section III-A-1 (with \( k = s \), as it should be.)

The general formula for the variance is quite lengthy. For \( b = 1 \),
\[
\text{Var}(w) = \frac{\left( 1 - \frac{1}{k} \right) p(1 - q)}{12(1 - p)^2} \cdot \left[ \left( 6 - 5 \left( 1 + \frac{1}{k} \right) p + 2 \left( 1 + \frac{1}{k} \right) p^2 \right) \right. \\
+ \left. \left( 6 - 5 \left( 1 + \frac{1}{k} \right) p + 2 \left( 1 + \frac{1}{k} \right) p^2 \right) q \right] \cdot \left( 5 - \frac{13}{k} \right) \cdot \left( 5 - \frac{13}{k} \right) p + 2 \left( 1 + \frac{1}{k} \right) p^2 \right) q^2 \\
+ \left( - \left( 5 - \frac{13}{k} \right) \right) p + 2 \left( 1 + \frac{1}{k} \right) p^2 \right) q^3 \\
+ \left. \left( 3 \left( 1 - \frac{1}{k} \right) - 6 \left( 1 - \frac{1}{k} \right) p \right) \right) q^3 .
\]
B. Geometric Service Distribution

Suppose that service times are geometrically distributed. Let the distribution of service times be 

\[ g_j = p(1 - p)^{j-1}, \]

\[ j = 1, 2, \ldots, \]

where \( p \) is a constant, \( 0 < p \leq 1 \). Then

\[ U(z) = \sum_{j=0}^{\infty} z^j \frac{\mu(1 - \mu)^{j-1}}{1-z(1-\mu)} = \frac{\mu z}{1 - z(1-\mu)} \]

so

\[ U'(1) = m = \frac{1}{\mu}, \]

\[ U''(1) = \frac{2(1-\mu)}{\mu^2}, \]

and

\[ U'''(1) = \frac{6(1-\mu)^2}{\mu^3}. \]

For the case of individual arrivals and balanced traffic, as in Section III-A-1 above, we have

\[ R(z) = \left( 1 - \frac{p}{s} + \frac{pz}{s} \right)^{k}, \]

\[ R'(1) = \lambda = \frac{kp}{s}, \]

\[ R''(1) = k(k-1) \left( \frac{p}{s} \right)^2 = \lambda^2 \left( 1 - \frac{1}{k} \right), \]

and

\[ R'''(1) = k(k-1)(k-2) \left( \frac{p}{s} \right)^3 = \lambda^3 \left( 1 - \frac{1}{k} \right) \left( 1 - \frac{2}{k} \right). \]

Recall that the traffic intensity is

\[ \rho = m \lambda = \frac{kp}{\mu s}. \]

Substituting into (2) and (3), we obtain

\[ \mathbb{E}(w) = \frac{\rho \left( 2 - \left( 1 + \frac{1}{k} \right) \mu \right)}{2 \mu (1-\rho)} \]

and

\[ \text{Var}(w) = \frac{6 \rho \left( 4 - 4 \mu \left( 1 + \frac{1}{2k} \right) + \mu^2 \left( 1 + \frac{1}{k} \right) \right) - \rho^2 \left( 12 - 12 \mu + 5 \mu^2 \left( 1 - \frac{1}{k^2} \right) \right) - 2 \rho \left( 6 - 12 \mu + 5 \mu^2 \left( 1 + \frac{1}{k^2} \right) \right)}{12 \mu^2 (1-\rho)^2}. \]

These reduce to the equations in Section III-A-1 when \( \mu = 1 \).

C. M/M/1 Queues

An M/M/1 queueing system can be thought of as the continuous time limit of systems with geometric service times. Suppose we scale time so that there are \( n \) cycles per time unit. Consider uniform traffic, single arrivals (as in Section III-B); let the geometric distribution approach the continuous, exponentially distributed, limit by replacing \( p \) by \( \mu n / s \), and letting \( n \) tend to infinity. (This gives the same average service time in unscaled units.) Replace the arrival probability \( p \) by \( p/n \) to maintain the same traffic intensity \( \rho \). Then as \( n \rightarrow \infty \), the queue becomes an M/M/1 queue with arrival rate \( \lambda = p k / s \) and service time \( m = 1/\mu \).

We can see this formally as a consequence of our results by computing the limiting distribution of the waiting time.

\[ R(z) = \left( 1 - \frac{p}{ns} + \frac{pz}{ns} \right)^{k}, \]

\[ = 1 + \frac{pk}{ns} (z-1) \]

and

\[ U(z) = \frac{z-1}{1 - z \left( 1 - \frac{\mu}{n} \right)}, \]

which yields

\[ t(z) = (1-\rho) \frac{1}{1 - \rho \left( 1 - \frac{\mu}{s} \right)} \]

Now scale time by replacing \( z \) by \( z^{1/n} \), and change the \( z \) transform to a Laplace transform by introducing the variable \( s = -\log z \). This gives the Laplace transform of waiting time as

\[ \frac{1 - \rho}{1 - \rho \left( 1 - \frac{\mu}{s} \right)} \]

which is the well-known transform of the M/M/1 waiting time distribution (see [9, Section 5.12]). So, in particular,

\[ \mathbb{E}(w) = \lim_{n \to \infty} \frac{1}{n} \left( 2 - \left( 1 + \frac{1}{k} \right) \mu \right) = \frac{1}{\mu} \frac{1 - \rho}{1 - \rho}. \]
This is the waiting time for an M/M/1 queue, as it should be. A similar derivation will show that the variance we obtained for geometric distribution converges to the variance of an M/M/1 queue.

D. Constant Service Times

We now consider the situation when messages can have one of several constant service times. We will only consider uniform traffic, single arrivals. Thus, as in Section III-A-1,

\[ R(z) = \left(1 - \frac{p + pz}{s}\right)^k, \]
\[ R'(1) = \lambda = \frac{kp}{s}, \]
\[ R^*(1) = k(k-1) \left(\frac{p}{s}\right)^2 = \lambda^2 \left(1 - \frac{1}{k}\right), \]
and
\[ R^{**}(1) = k(k-1)(k-2) \left(\frac{p}{s}\right)^3 = \lambda^3 \left(1 - \frac{1}{k}\right) \left(1 - \frac{2}{k}\right). \]

1) Single Size: First, suppose that each message takes exactly \( m \) units of time to transmit. This will occur, for instance, when each message is composed of an equal number \((m)\) of packets, and the constituent packets of a message are transmitted at consecutive cycles. Then

\[ U(z) = z^m, \]
so that
\[ U'(1) = m, \]
\[ U^*(1) = m(m-1), \]
and
\[ U^{**}(1) = m(m-1)(m-2). \]

The traffic intensity is now

\[ \rho = mpk. \]

Substituting into (2) and (3), we obtain

\[ E(w) = \frac{\rho \left(\frac{m-1}{k}\right)}{2(1-\rho)} \]

and

\[ \text{Var}(w) = \frac{\rho \left[6m - 5\rho \left(1 + \frac{1}{k}\right) + 2\rho^2 \left(1 + \frac{1}{k}\right)^2\right] + (m-1)[2(2m-1)-\rho(m+1)]}{12(1-\rho)^2}. \]

These coincide, for \( m = 1 \), with the equations of Section III-A-1.

2) Multiple Sizes: Now suppose there are \( n \) service times \( m_1, \cdot \cdot \cdot, m_n \), and service time \( m_i \) occurs with probability \( g_i \). This will occur when there are different kinds of requests. For example, read requests are likely to have different sizes than write requests.

We get

\[ U(z) = \sum_{i=1}^{n} g_i z^{m_i}, \]
so that
\[ U'(1) = \sum_{i=1}^{n} m_i g_i, \]
and
\[ U^*(1) = \sum_{i=1}^{n} m_i(m_i-1) g_i. \]

Thus,
\[ \rho = \frac{kp}{s} \sum_{i=1}^{n} m_i g_i. \]

Substituting into (2), we obtain

\[ E(w) = \frac{\lambda \sum_{i=1}^{n} m_i \left(\frac{m_i-1}{k}\right) g_i - \rho \sum_{i=1}^{n} m_i \left(\frac{m_i-1}{k}\right) g_i}{2(1-\rho) \sum_{i=1}^{n} m_i g_i}. \]

The formula for the variance could also be obtained, but it is quite lengthy and not particularly enlightening.

IV. LATER STAGES

We do not know how to analyze the later stages exactly as the inputs at successive cycles are not independent. We have, however, developed some very useful approximate formulas for the average and variance of the waiting time. These are based on two observations. First, as we progress through the network, the waiting time statistics quickly approach a limiting distribution. Second, nearly every waiting time distribution in queueing theory has an average on the order of \( 1/(1-\rho) \) as \( \rho \) tends to one; that is, if \( w_i(\rho) \) is the average waiting time at the \( i \)th stage, and \( w_i(\rho) \) is the limit as \( i \) gets large, we expect \( \lim_{i \to \infty} (1-\rho) w_i (\rho) \) to exist. We calculated \( w_i(\rho) \) exactly in Section III, and we expect \( w_i(\rho) \) and \( w_i(\rho) \) to have similar qualitative properties, i.e., they should depend on most
parameters in roughly the same way. Hence, it seems reasonable to estimate
\[ r(p) \equiv \frac{w_p}{w_1}. \]

It seems clear that, for uniform traffic with unit service time, \( r(0) = 1 \). We use simulations to estimate \( r(1/2) \), and then simply linearly interpolate to obtain a value \( a \) such that
\[ r(p) \approx 1 + ap. \]  
(10)

Then
\[ w_p = (1 + ap)w_1. \]

We will also generalize the formulas to take into account the dependence of \( w_p \) on the stage, the switch size, and the message size distribution. This method of interpolation was previously applied to queueing systems by Burman and Smith [3] using light and heavy traffic theory. The light traffic limit exists in our case. We do not have a heavy traffic analysis for our process, so we rely on simulation instead.

Using the same ideas, we can obtain an approximation for the variance. Let \( v_i \) be the variance of the waiting time at stage \( i \), and \( v_m \) be the variance at the stage. Since the formulas for variance have one higher power of \( p \), we expect a good approximation of \( v_m/v_1 \) to contain (at least) one higher power of \( p \), i.e., we obtain a quadratic interpolation for the variance. The variance after several stages can be approximated by
\[ v_m = (1 + ap + bp^2)v_1, \]
where \( a \) and \( b \) are constants to be determined.

In the remainder of this section, we obtain our expression for the waiting time and variance in step-by-step generalizations. We first estimate them for uniform traffic when messages have size one, then size \( m \), and then general size. Finally, we consider nonuniform traffic.

**A. Service Time One**

Consider service time one, \( 2 \times 2 \) switches, single arrivals, and uniform traffic. For \( p = 0.5 \), \( w_1 = 0.25 \) [see (6)], and, from the simulations in Table I, \( w_m \) seems to be about 0.3. Substituting into (10) and solving for \( a \) gives \( a \approx 2/5 \). Thus, we find
\[ r(p) = 1 + 2p/5, \]
so the waiting time
\[ w_m = \left( 1 + \frac{2p}{5} \right) \frac{p}{4(1 - p)}. \]

Table I compares the simulation results to our formulas. The waiting time values in the ANALYSIS row are from the exact formula for the first stage [see (6) and (7)], and the waiting time values in the ESTIMATE row are from the above approximation for the waiting time in the limit. Note that the approximation seems to be slightly low for \( p \) small and slightly high for \( p \) large. More complete simulation results (not included for brevity) show that \( r(p) \) is actually slightly concave. An even better estimate could be obtained by using a quadratic approximation.

Using the same technique for \( 4 \times 4 \) and \( 8 \times 8 \) switches gives \( a \) a bit less than 0.2 and \( a \) a bit less than 0.1, respectively (see (6) and Table II). This suggests that the above formula can be (crudely) extended to \( k \times k \) switches by linearly including \( k \) as a parameter. This gives the waiting time
\[ w_m = \left( 1 + \frac{4p}{5k} \right) \frac{1 - k}{p} \frac{1}{2(1 - p)}. \]  
(11)

Table II compares the simulation results to our formulas.

In Tables I and II it looks as if \( w_m \) approaches \( w_m \) geometrically. This suggests a formula of the form \( r(p) = 1 + (1 - \alpha^{i+1})(r(p) - 1) \) for some \( \alpha < 1 \), yielding
\[ w_i = \left( 1 + \frac{4p}{5k} \frac{1 - \alpha^{i+1}}{1 - \alpha} \right) \frac{1 - k}{p} \frac{1}{2(1 - p)}. \]  
(12)

as the expected waiting time at the \( i \)th stage. Looking once again at the formula for \( k = 2 \) and \( p = 0.5 \) (Table I), gives \( \alpha = 2/5 \) as a good approximation. It turns out that this value of \( \alpha \) works reasonably well for general \( k \) and \( p \). For brevity, we do not explicitly compare this formula to the simulations (although the interested reader can easily do the calculations).

It is by no means surprising that, for a given \( p \) and \( k \), \( w_i \) approaches \( w_m \) geometrically; what is perhaps surprising is that a single value of \( \alpha \) works well for all \( p \) and \( k \). Applying the same techniques to variance, we find that a reasonable formula for the variance after several stages is
\[ v_m = \left( 1 + \frac{p}{2k} + \frac{p^2}{k} \right) \left[ \frac{1 - \frac{1}{k}}{p} \left( 6 - 5p \left( 1 + \frac{1}{k} \right) + 2p^2 \left( 1 + \frac{1}{k} \right) \right) \right] \frac{1}{12(1 - p)^2}. \]  
(13)
Since this is only an approximation and since the simulation results do not give exact answers, there is quite a bit of freedom in choosing coefficients \(a\) and \(b\) for the \(p\) and \(p^2\) terms. Other choices will surely work just as well or better.

We can also estimate the variance at stage \(i\) to be

\[
v_i = \left(1 + \left(\frac{p}{2k} + \frac{2p^2}{k}\right)(1-\alpha^{i-1})\right) \cdot \frac{1}{12(1-p)^2} \left[6 - 5p \left(\frac{1}{1-k} + 2p \left(\frac{1}{1-k}\right)\right)\right]
\]

where \(\alpha = \frac{2}{5}\).

### B. Single Service Time

Consider the case when messages have a single constant size. Our model of the first stage is not a particularly good model for the later stages. At the first stage, a source after sending a message can send a new message on the next or any later cycle. At later stages, since sources are outputs from queues, once a source sends a message, that source will not send a message for at least \(m\) cycles. This will tend to reduce queuing delays at the later stages.

Later stages can be better modeled by assuming that messages take one cycle to be processed, but the cycle time is \(m\) times as long. Following [11], we use the formula for service time one (11), and, for a fixed \(p\), multiply the time to process a message by a factor \(m\), and also multiply the average number of packets per cycle by \(m\). In other words, for a fixed traffic intensity \(\rho\), the cycle time is \(m\) times as large. This gives the average waiting time

\[
w_m \approx \left(1 + \frac{4mp}{5k}\right) \frac{(1-\frac{1}{k})m^2p}{2(1-mp)}.
\]

For \(m \geq 2\), this formula is a reasonable approximation at all stages after the first, and, of course, we have an exact formula for the first stage. Table III compares the simulation results to our formulas.

Let us examine the behavior of the interior stages in light traffic. Reasoning as in Section III-C, if we allow \(m\) to increase and \(p\) to decrease with \(mp = \rho\) constant, then in time scaled by \(m\), the first stage output queues become M/D/1 queues with arrival rate \(p\) and service time 1. (Actually, the well-known waiting time statistics of M/D/1 queues can be obtained as limits of (8) and (9).) Now the interior stages are not precisely M/D/1 queues in this scaling, because the packets output from previous stages must be spaced by at least \(m\) time units. Nevertheless, it is clear that in light traffic the interior stages will resemble M/D/1 queues, but the congestion will be lower than at the first stage, since packets will be very unlikely to collide with other packets from the same source. That is, the congestion will be as if the arrival rate were \((1-\frac{1}{k})p\). Using the M/D/1 light traffic results,

\[
E(w) = \frac{\left(\frac{1}{k}-1\right)\rho}{2} + O(\rho^2)
\]

and

\[
\text{Var}(w) = \frac{\left(\frac{1}{k}-1\right)\rho}{3} + O(\rho^2).
\]

Our approximations should have these properties, too. Equation (15) does satisfy this.

To obtain an approximation for the variance, we argue as before: start with the formula for the variance at the first stage for unit size messages (7), multiply by \(m^2\), change \(p\) to \(mp\), and then use the light traffic analysis and the simulations to interpolate. Our heuristic formula is [see (13)]

\[
v_m = \left(\frac{2}{3} + \frac{C_1\rho}{k} + \frac{C_2\rho^2}{k}\right) \cdot \left[6 - 5mp \left(\frac{1}{1-k} + 2(mp)^2 \left(\frac{1}{1-k}\right)\right)\right] \frac{1}{12(1-mp)^2},
\]

where \(2/3\) was obtained from light traffic analysis. Light traffic analysis is a limiting case for \(m\) large; in practice, we found that 7/10 works better than 2/3 for small and moderate message sizes. We match the constants \(C_1\) and \(C_2\) to simulation results, giving

\[
v_m = \left(\frac{2}{10} + \frac{3mp}{10k} + \frac{4(mp)^2}{10k}\right) \cdot \left[6 - 5mp \left(\frac{1}{1-k} + 2(mp)^2 \left(\frac{1}{1-k}\right)\right)\right] \frac{1}{12(1-mp)^2}.
\]

This approximation is still slightly low for \(m\) small, as can be seen in Table III. Better approximations can be obtained for each individual value of \(m\); in particular, (13) is a much better approximation for \(m = 1\). As with waiting times, for \(m \geq 2\), this formula can be used to approximate variances for all stages after the first.

### C. Multiple Service Times

As in Section III-D-2, suppose there are \(n\) service times \(m_1, \ldots, m_n\), and service time \(m_i\) occurs with probability \(\xi_i\).
The average service time is $m = \sum_{i=1}^{n} g_i m_i$. To obtain an approximate formula for the average waiting time, replace the size of all messages by their average size $m$ and use the approximate waiting time formula from the previous section (15). This gives the average waiting time

$$w_m = \frac{m^2 p}{2(1-mp)}.$$

The values obtained from this formula tend to be a bit low. The reason is that we are approximating multiple size messages by their average size. Since we are able to calculate everything at the first stage exactly, we know how much off such an assumption would be at the first stage: simply the ratio of the actual expected waiting time and the waiting assuming all messages have their average size. Assuming this ratio is fairly constant at the different stages, multiplying the above formula by this ratio gives a very good approximation:

$$w_m = \frac{1 + 4mp}{sk} \frac{m^2 p}{2(1-mp)}.$$

An approximate formula for the variance $v_m$ could be obtained similarly, but, as with the variance formula for the first stage, it is quite lengthy. We have, however, obtained numerical values from both variance formulas, i.e., for $v_t$ and $v_m$. Table IV compares the simulation results to our formulas.

### D. Nonuniform Traffic

We can also obtain approximate formulas when each input has a distinct favorite output port. Our form for $w_m$ is a linear function of $q$ multiplied by the exact formula for the first stage. For $m = 1$, the coefficients of the linear function were found by starting with (11) ($q = 0$ case) and comparing to the simulation results. The average waiting time in the limit can be approximated as

$$w_m = \left(1 + \frac{4(1-q)p}{5k}\right) \frac{(1-q^2) (1-\frac{1}{k})p}{2(1-p)}.$$

Similarly, the variance $v_m$ can be approximated as a linear function of $q$ multiplied by the exact formula for the variance at the first stage. Starting with (13) and comparing to simulation results gives

$$v_m = \left(1 + \frac{p}{2k^2 + p^2} \right) \frac{(1-q)}{(1-\frac{1}{k})} \left(\frac{1}{k} p \left[6-5p \left(1+\frac{1}{k} \right) + 2p^2 \left(1+\frac{1}{k} \right) \right] \right)/12(1-p)^2.$$
of the waiting time at each stage, these can be used to obtain approximations for the total waiting time. The expected value of the total waiting time is simply the sum of the average messages of size one, summing the waiting times at the different stages. In particular, for $0.5$, the total variance for an $n$-stage network as the variances should be a good approximation. In particular, geometrically as stages become further apart. Thus, summing have fairly low correlation, and the correlation seems $n(l+(r+T)(l--))$

For messages of size one, summing the correlation of waiting times between stages for the variance of the total waiting time would simply be the sum is the case with Poisson arrivals and exponential service times, of the variances at the different stages. Table VI shows the can be approximated as the average waiting time from the first stage [see (2)]

$$w_{m} \approx \left(1 + \frac{4mp}{5k}\right) \left(1 - q^{2}\right) \left(1 - \frac{1}{k}\right) \frac{m^{2}p}{2(1-mp)}.$$  
This is a good approximation for the average waiting time at all stages after the first ($m \geq 2$). The variance can be approximated similarly.

V. TOTAL DELAY

Once we have formulas for the expected value and variance of the waiting time at each stage, these can be used to obtain approximations for the total waiting time. The expected value of the total waiting time is simply the sum of the average waiting times at the different stages. In particular, for messages of size one, summing the $w_{i}$ in (12) approximates the total waiting time for an $n$-stage network as

$$n \left(1 + \frac{4p}{5k} \left(1 - \frac{1 - \alpha^{n}}{n(1-\alpha)}\right) \right) \left(1 - \frac{1}{k}\right) \frac{m^{2}p}{2(1-mp)}.$$  
where $\alpha = 2/5$. For $m > 1$, the average total waiting time can be approximated as the average waiting time from the first stage [see (8)] plus $n - 1$ times the waiting time as the later stages [see (15)], which is

$$\frac{(m-1)mp}{2(1-mp)} + (n-1) \left(1 + \frac{4pm}{5k}\right) \left(1 - \frac{1}{k}\right) \frac{m^{2}p}{2(1-mp)}.$$  
If the waiting times from stage to stage were independent, as is the case with Poisson arrivals and exponential service times, the variance of the total waiting time would simply be the sum of the variances at the different stages. Table VI shows the correlation of waiting times between stages for $k = 2$, $p = 0.5$, and $m = 1$. Note that waiting times at neighboring stages have fairly low correlation, and the correlation seems to drop geometrically as stages become further apart. Thus, summing the variances should be a good approximation. In particular, for messages of size one, summing the $v_{ij}$ in (14) approximates the total variance for an $n$-stage network as

$$\sum_{j=1}^{n} \left(1 + \frac{2a(1-b^{*}-1)}{1-b} \right) v_{ij}.$$  
For $m = 1$, we use the $v_{i}$ from (14). For $m > 1$, $v_{i}$ is the true variance for the first stage [see (9)], and $v_{i}$, $i > 1$, can be approximated by the formula for $v_{m}$ [see (16)]. Tables VII–XII compare the simulation results to our formulas.

The distribution of waiting times seems to be about the same for all stages. If the distributions were independent, by the central limit theorem, the total waiting times for a large number of stages could be approximated by a (truncated) normal distribution, whose sum is the sum of the expected values and whose variance is the sum of the variances. The central limit theorem actually holds under much weaker hypotheses than independence (see, for example, [2]), and we expect it essentially to apply here. For only a few stages, however, a normal approximation may not be very accurate at the tails. Typically in queueing systems, the distribution of waiting times has an exponential or geometric tail, so we expect a gamma distribution with the proper expected value and variance to be a good approximation for even small networks.

We have simulated networks of $n = 3, 6, 9, 12$ stages for service times $m = 1, 4$ and traffic intensities $\rho = 0.2, 0.5, 0.8$. The histograms in Figs. 3–8 show for each simulation the probability that a message has a given total waiting time. (Note that a message can have waiting time zero, since the delay does not include the service time of the message itself.) For each figure, the smooth curve is the gamma distribution with the expected value and variance as estimated in Tables VII–XII. The figures show an incredibly good match between the estimated and the observed distributions, especially at the tails. In practice, the expected value, the variance, and the tail of the waiting time distribution are the quantities of interest; we believe our formulas are accurate enough for all practical purposes.
In order to obtain the total time is one, then the total service time is simply the number of successive stages. In general, it is the sum of service times at the previous stage. However, the correlation is weak, so that these random variables are stochastically nearly independent. Thus, the variance of the total delay is approximately the variance of the total waiting time plus the sum of the variances of the service times. If the service times are constant, then their variances are zero, so the variance of the total delay is exactly the variance of the total waiting time. In general, the distribution of the total delay can easily be approximated by looking at the distributions of individual service times and the distribution of the total waiting time.

Note that constant service time greater than one was used to model a network in which messages are transmitted in several consecutive packets. In such situations, waiting (or service) at one stage can start before service at the previous stage has terminated; the total service time in the network is \( n + m - 1 \), where \( n \) is the number of stages, and \( m \) is the number of packets per message.

### VI. Conclusion

We have analyzed the delay experienced by a message in a buffered, multistage, packet-switching banyan network. For the first stage, we were able to derive the complete distribution of the delay for a very general class of distributions, assuming messages have discrete sizes. This was used to determine exactly the average and variance of the delay for several commonly considered distributions. The formulas can be easily applied to other distributions. Using the delay formulas for the first stage, we developed extremely good approximations for the average and variance of the delay at later stages. Finally, this allowed us to obtain good approximations for the full distribution of the total delay of a message through the entire network.

In order to approximate the delay after the first stage, it was essential to have good formulas for the delay at the first stage. It was only by building on them that we were able to make educated guesses as to the delays at later stages.

One aspect of our results that is worth stressing is the dependency of waiting time on the message size \( m \). For a fixed traffic intensity \( p \), the average waiting time increases linearly in \( m \) [see (8) and (15)], and the variance increases quadratically in \( m \) [see (9) and (16)]. Thus, while using larger messages may save the overhead of duplicating the same routing information over several packets, it may dramatically increase delays in all but very lightly loaded networks. This point has already been made in [11], [12], and [8], but does not seem to be widely appreciated.

We believe that our techniques for obtaining approximations can be usefully extended in several ways. For example, better and more extensive simulations would allow the approximate formulas to be improved. Although it may not be tractable to analyze the delay at later stages exactly or even asymptotically, it might be possible to obtain a heavy traffic analysis. This would provide an exact value for \( \lim_{\rho \to 1} r(\rho) \), and would simplify the task of obtaining good approximations for \( w_m \) and \( v_m \). Given our formulas for infinite buffer delays, along with some simulation results for finite buffers, it is possible that one could develop good approximate formulas for finite buffer delays.

Finally, while our simulations seem to indicate clearly that average waiting time at successive stages converges, it would be nice to be able to prove this result formally, i.e., to show...
Fig. 3. Distribution of waiting times—simulation and prediction.

Fig. 4. Distribution of waiting times—simulation and prediction.
Fig. 5. Distribution of waiting times—simulation and prediction.

Fig. 6. Distribution of waiting times—simulation and prediction.
Fig. 7. Distribution of waiting times—simulation and prediction.

Fig. 8. Distribution of waiting times—simulation and prediction.
that average delay at successive stages can be bounded, independently of the network size.

ACKNOWLEDGMENT

The authors thank S. Abraham, I. Mitrani, and B. Muchmore for their help.

REFERENCES


Marc Snir received the B.Sc. degree in 1972 and the Ph.D. degree in 1979, both in mathematics, from the Hebrew University of Jerusalem, Jerusalem, Israel. He is currently at the IBM T. J. Watson Research Center, on leave from the Hebrew University of Jerusalem. From 1979 to 1980 he held a visiting position at the University of Edinburgh. From 1980 to 1982 he was Assistant Professor at New York University, where he worked on the design of the NYU Ultracomputer, a large-scale parallel computer. Since 1982 he has been a Senior Lecturer in the Department of Computer Science of the Hebrew University. In addition to interconnection networks, his current research interests include parallel computer architectures, parallel algorithms, resource management in parallel computers, parallel programming languages, and complexity theory.

Dr. Snir is a member of IEEE Computer Society and of the Association for Computing Machinery.

Alan Weiss was born on December 5, 1955. He received the B.S. degree in both mathematics and physics from Case Western Reserve University, Cleveland, OH, in 1976, and the M.S. and Ph.D. degrees in mathematics from the Courant Institute of New York University, New York, NY, in 1979 and 1981, respectively.

Since 1981 he has been a member of the Technical Staff in the Mathematical Sciences Research Center at AT&T Bell Labs, Murray Hill, NJ. His research interests include the theory of large deviations, stochastic processes, and equations of evolution.

Clyde P. Kruskal was born on May 25, 1954. He received the A.B. degree in mathematics and computer science from Brandeis University, Wal-tam, MA, in 1976, and the M.S. and Ph.D. degrees in computer science from the Courant Institute of New York University, New York, NY, in 1978 and 1981, respectively.

He was an Assistant Professor in Computer Science at the University of Illinois, Urbana, from 1981 to 1985. Since 1985, he has been on the faculty at the University of Maryland, College Park and is now an Associate Professor. His research interests include the analysis of sequential and parallel algorithms and the design of parallel computers.