An Analysis of Banyan Networks Offered Traffic With Geometrically Distributed Message Lengths

I-Pyen Lyaw and David M. Koppelman*
Louisiana State University, Baton Rouge, LA 70803

Abstract: An analysis of finite-input-buffered banyan networks offered traffic having geometrically distributed message lengths is presented. This is one of the few multistage-network analyses for networks offered non-unit-length messages and is the only one that the authors are aware of for finite-input-buffered banyan networks. In the analysis, network switching elements are modeled using two state-machines, one for queue heads (HOL's), the other for entire queues. A network is modeled using one switching-element model to represent each stage. Together these model the effect that non-unit-length messages have on banyans. Solutions are obtained iteratively. Network performance figures were obtained with this analysis and compared to simulation results. The figures show that the analysis can predict the effect of message length on throughput and delay, including the performance degradation caused by longer messages.

* This work is supported in part by the Louisiana Board of Regents through the Louisiana Education Quality Support Fund, contract number LEQSF (1993-95)-RD-A-07 and by the National Science Foundation under Grant No. MIP-9410435.

1 INTRODUCTION

Banyan networks, unique-path multistage interconnection networks [1], are widely considered for use in communication and parallel-computing systems. Performance analyses of such networks are needed both for evaluation of system designs and for understanding the networks themselves. Many banyan-network analysis methods have been reported; the bulk of the work was for networks offered unit-length (single-packet) messages. For example, Patel analyzed unbuffered banyans [8], Jenq, single-buffered banyans [3], Yoon, Lee, and Liu, finite-buffered banyans [9], and Mun and Youn described a finite-buffered banyan analysis which works well at heavy traffic loads [7]. (All of these analyses were for input-buffered networks.) As is readily observed from simulation, message length has a strong effect on performance. Since networks used for communication switches and parallel computers must carry messages having varying lengths, a non-uniform-message-length analysis is needed.

Such an analysis had been performed by Kruskal, Snir, and Weiss [4] for infinite-buffered networks. The networks they analyzed have output-buffered switching elements (SE's) in which queues can be simultaneously fed by any number of SE inputs. Exact first-stage switching-element queue state distributions were found. An empirically derived formula was then used to find the waiting time in subsequent stages. Because messages entering a queue are not blocked, a wide variety of offered-traffic models can be analyzed, including those with geometrically distributed message lengths. Their analysis of such traffic shows that delay increases as average message length is increased (while holding traffic intensity constant) [5]. Their analysis, however, is not applicable to many of the networks that might actually be used. Actual networks use finite queues and may also use crossbars that block. These create a *back pressure* effect [6] which results in a distribution of messages within the network very different than the networks Kruskal *et al* analyze. The modeling of message distribution is an important part of analysis, and so a different model must be used for finite-buffered, blocking-crossbar networks.

The model used here captures the following behavior of non-unit-length messages in these networks. A message, while passing from a queue in one SE to another, will be the only message using its SE output. Message packets following the first packet enter the next-stage SE at, what amounts to, an arrival rate of one, tending to fill the queue there. Thus, the probability that the first packet of a message (the *head packet*) will find the next-stage queue full is lower than the corresponding probability for other packets in the message.

To capture this behavior, each stage is modeled by two state machines. One for the queue heads in a switching element, the *head-of-line (HOL) model,* the other for a single (entire) queue, the *queue model.* (Combined HOL/queue SE models have been used earlier, for example by Hui [2] to analyze networks with unit-length messages. The HOL and queue models here are different.) The HOL model encodes the state of the queues' head slots (whether the slot is empty or has a head or non-head packet, as well as its destination). It is used to find the distribution of these states, from which arrival and service rates for the queue models are computed. The queue model has two sets of states: one set is for queues into which a message is entering (that is, the head packet of the message has entered but the last packet, the *[tail packet],* has not yet entered), and one set of states is for queues into which no message is entering. Transitions for the HOL model are based, among other things, on the probability that a message using a switching-element output will end. Transitions are also based on the probability that there will be space in the next-stage queue given the type of packet, head or non-head. These space probabilities are determined from the queue model. These model the behavior described above.

The remainder of this paper is organized as follows. In the next section network-structure and other preliminaries appear. The analysis is described in Section 3 [p. 4], the analysis is compared with simulation in Section 4 [p. 13]. Conclusions follow in Section 5 [p. 15].

2 PRELIMINARIES

2.1 NETWORK STRUCTURE

Analyzed networks will be specified by a 3-tuple, (n, a, m). Such a network consists of

n stages, numbered 1, the *input stage,* to n, the *output stage.* Each stage contains a^{n-1} a-input, a-output *switching elements (SE's).* Each SE consists of a m-slot queues, each connected to the crossbar inputs. Links connect SE's in adjacent stages and first- and last-stage SE's to network inputs and outputs, respectively. The links can be connected in any pattern for which there is exactly one path between all network input /output pairs. See [1,2,6] for details.

2.2 Message Structure and Flow Control

Data arrives at network inputs in the form of *messages.* Each message consists of a number of fixed-length *packets* which pass through the network as a unit. The first packet is called the *head packet* and the last packet is called the *tail packet;* symbols H and T are used to refer to head packets and tail packets, respectively. Symbols $\overline{\mathsf{H}}$ and $\overline{\mathsf{T}}$ refer to packets which follow the head packet and precede the tail packet, respectively. Switching-element queues use a first-in, first-out, service discipline. Each queue slot holds exactly one packet. The slot that holds, or would hold, the next packet to leave is called the *head-of-line (HOL)* slot. A packet occupying the HOL slot is called the *HOL packet.* Each message has a *destination,* the network output to which it is bound. A HOL packet's *needed link* is the link (connected to the SE output) on the path to the packet's destination. A HOL packet's *next-stage queue* is the next-stage queue that is on the path to the packet's destination.

Time is divided into *cycles.* A HOL packet will move to the next-stage queue (or network output) during a cycle if there is space and it wins contention for the needed crossbar output. There is always space at a network output. There is space in a queue if there is at least one slot free during the cycle. Any queue space vacated will be available at the next cycle. A non-head packet will always win contention. Otherwise, for each SE output link, one winner will be randomly chosen from those HOL packets needing the link.

2.3 Traffic Model

The statistics of messages arriving at the network inputs are independent and identically

distributed. During a cycle a network input can either be idle, have a head packet arriving, or have a non-head packet arriving. Message arrival times and lengths are described by a three-state discrete-time Markov chain, with states labeled I, (idle); S, (start); and A, (active). Let $x \rightarrow y$ denote a transition from state x to y and let $\tau_{x \rightarrow y}$ denote the corresponding transition probability. An input not receiving a packet during a cycle is modeled by transitions $x \rightarrow \mathsf{I}$, $x \in \{\mathsf{I}, \mathsf{S}, \mathsf{A}\}$. A head packet arrival is modeled by transitions $x \rightarrow \mathsf{S}$, $x \in \{\mathsf{I}, \mathsf{S}, \mathsf{A}\}$; a message continuing is modeled by transitions $x \rightarrow \mathsf{A}$, $x \in \{\mathsf{S}, \mathsf{A}\}$. Transition probabilities are $\tau_{\mathsf{I} \rightarrow \mathsf{S}} = \lambda$, $\tau_{\mathsf{I} \rightarrow \mathsf{I}} = 1 - \lambda$, $\tau_{\mathsf{S} \rightarrow \mathsf{A}} = \tau_{\mathsf{A} \rightarrow \mathsf{A}} = 1 - \mu$, $\tau_{\mathsf{A} \rightarrow \mathsf{S}} = \tau_{\mathsf{S} \rightarrow \mathsf{S}} = \mu\lambda$, and $\tau_{\mathsf{A} \rightarrow \mathsf{I}} = \tau_{\mathsf{S} \rightarrow \mathsf{I}} = \mu(1 - \lambda)$. This traffic model generates messages with an expected length of $1/\mu$ packets and a flow rate of $(\mu/\lambda - \mu + 1)^{-1}$ packets per cycle, where $\lambda, \mu \in (0, 1]$.

Message destinations are randomly chosen (once for each message) and uniformly distributed over all outputs. As a consequence of this destination distribution and the banyan network's structure, the link needed by a head packet entering a HOL slot will be uniformly distributed over the SE outputs.

3 ANALYSIS

3.1 OVERVIEW

A network is modeled by n Markov-chain pairs, each of which characterizes a stage. One pair member, called the *queue model,* characterizes a switching-element queue. The other, called the *HOL model,* characterizes a switching-element HOL system. All queue and HOL models making up a network are statistically independent of each other. The state distributions are solved by iteratively computing state distributions and transition probabilities.

The HOL model is important because messages are transmitted in several consecutive packets which cannot be interrupted. If the correlation among packets by prior contention, captured in the HOL model, is instead neglected, predictions will be inaccurate, especially for large message lengths.

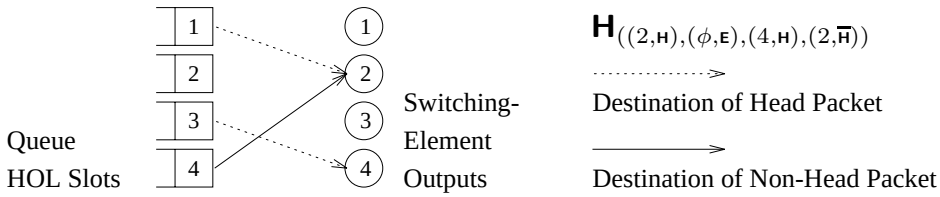

Figure 1. HOL State Example.

3.2 HOL MODEL

HOL-model states are labeled $\mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}$, where $d_i \in \{1,2,\ldots,a,\phi\}$ and $l_i \in \{\mathsf{H},\overline{\mathsf{H}},\mathsf{E}\}$ for $1 \leq i \leq a$. Pair (d_i,l_i) indicates the state of the HOL slot in SE queue i. Slot contents is indicated by l_i; H denotes a head packet, $\overline{\mathsf{H}}$ denotes a non-head packet, and E denotes an empty queue. The switching-element output needed by the HOL slot is indicated by $d_i \in \{1,2,\ldots,a,\phi\}$, where an integer refers to a switching-element output and ϕ denotes an empty queue. A HOL-model state is made up only of pairs $\{(\phi,\mathsf{E})\} \cup \{(d,l) \mid d \in \{1,2,\ldots,a\}, l \in \{\mathsf{H},\overline{\mathsf{H}}\}\}$. An example of a HOL-model state is illustrated in Figure 1.

The number of HOL states is large, even for systems with small switching elements. For purposes of analysis, the set of HOL states can be partitioned into a much smaller set of equivalence classes such that only one member of each class need be considered. See the appendix for details.

The probability that a stage-j HOL model is in state $\mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}$ is denoted $p_j(\mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))})$. Let H_α and H_β be two equivalent states. State transitions will be defined so that $p_j(H_\alpha) = p_j(H_\beta)$ for $1 \leq j \leq n$, as will be explained below.

HOL-model transition probabilities will be specified as a product of *HOL factors.* For each HOL-model transition probability there are a HOL factors, one for each of the queues in a SE. Let the HOL factor associated with queue i in stage j for a transition from H_t to H_{t+1} be denoted $f_j(H_t,H_{t+1},i)$ and let the HOL-model transition probability be denoted $T_j(H_t,H_{t+1})$. Then,

$$T_j(H_t,H_{t+1}) = \prod_{i=1}^{a} f_j(H_t,H_{t+1},i). \tag{1}$$

Let $p_j(H_t)$ be the probability of stage-j HOL-model state H at time t. Then the state probability

is defined to be

$$p_j(H_{t+1}) = \sum_{H_t \in \mathcal{H}} p_j(H_t) T_j(H_t, H_{t+1}),$$

for all $H_{t+1} \in \mathcal{H}$, where $\mathcal{H}$ is the set of all HOL-model states.

The value for a HOL factor is determined by the queue's role in the transition. A queue is said to be *active* in a transition from state H_t to H_{t+1} if it contains at least one packet in H_t and could have won the contention or retains control of the port in the transition to H_{t+1}. (Note that it is not always possible to determine if a queue wins contention in a transition between HOL-model states.) Define $A(H_t, H_{t+1}, i)$ to be true if queue i is active and false otherwise. Then A is given by the logical expression

$$A(\mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}, \mathbf{H}_{((D_1,L_1),(D_2,L_2),\ldots,(D_a,L_a))}, i) = \tag{2}$$
$$(l_i \neq \mathsf{E}) \wedge \forall_{x \in \{1,2,\ldots,i-1,i+1,\ldots,a\}} (d_x = d_i) \Rightarrow [(l_x = L_x = \mathsf{H}) \wedge (d_x = D_x)]$$

where $\wedge$ is the conjunction operator, $x \Rightarrow y$ is the logical implication operator, and $\forall$ is the universal quantifier. (*E.g.*, expression $\forall_{0<i<9}\, x_i$ is equivalent to $x_1 \wedge x_2 \wedge \cdots \wedge x_8$.)

HOL factors for active queues are determined by the probability of space in the next stage, the probabilities that a HOL slot contains a head packet, non-head packet, or is empty, and the number of queues needing the same port. A queue which is not active is either empty or did not win contention. The HOL factor for the former case is based on the probability of an arrival to the queue, the HOL factor for the later case is 1. HOL-factor expressions will be given after arrival and space probabilities are introduced.

Define $\sigma_{\mathsf{H},j}$ to be the probability that a head packet in the HOL slot of an active stage-j queue will find space in its next-stage queue. Similarly, define $\sigma_{\overline{\mathsf{H}},j}$ to be the probability that a non-head packet in the HOL slot of an active stage-j queue will find space in its next-stage queue.

Define $v_{\mathsf{E},j}$ to be the probability of head-packet arrival to an empty stage-j queue. Similarly, define $v_{\mathsf{T},j}$ to be the probability of head-packet arrival to a stage-j queue HOL slot given that the slot held a tail packet in the previous cycle.

The queue-i HOL factor for a transition from state $\mathsf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}$ to state $\mathsf{H}_{((D_1,L_1),(D_2,L_2),\ldots,(D_a,L_a))}$ is given by

$$f_j(\mathsf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}, \mathsf{H}_{((D_1,L_1),(D_2,L_2),\ldots,(D_a,L_a))}, i) = f_j(H_t, H_{t+1}, i) =$$

$$\begin{cases} (1 - v_{\mathsf{E},j}), & \text{if } L_i = l_i = \mathsf{E}; \\ \frac{\sigma_{\mathsf{H},j}(d_i)}{C(H_t,i)}\mu(1 - v_{\mathsf{T},j}), & \text{if } L_i = \mathsf{E},\ l_i = \mathsf{H},\ A(H_t, H_{t+1}, i); \\ \sigma_{\overline{\mathsf{H}},j}(d_i)\,\mu(1 - v_{\mathsf{T},j}), & \text{if } L_i = \mathsf{E},\ l_i = \overline{\mathsf{H}},\ A(H_t, H_{t+1}, i); \\ \frac{v_{\mathsf{E},j}}{a}, & \text{if } L_i = \mathsf{H},\ l_i = \mathsf{E}; \\ 1, & \text{if } L_i = l_i = \mathsf{H},\ D_i = d_i,\ \overline{A(H_t, H_{t+1}, i)}; \\ \left[(1 - \sigma_{\mathsf{H},j}(d_i)) + \sigma_{\mathsf{H},j}(d_i)\,\mu\frac{v_{\mathsf{T},j}}{a}\right]^{\frac{1}{C(H_t,i)}}, & \text{if } L_i = l_i = \mathsf{H},\ D_i = d_i,\ A(H_t, H_{t+1}, i); \\ \frac{\sigma_{\mathsf{H},j}(d_i)}{C(H_t,i)}\mu\frac{v_{\mathsf{T},j}}{a}, & \text{if } L_i = l_i = \mathsf{H},\ D_i \neq d_i,\ A(H_t, H_{t+1}, i); \\ \sigma_{\overline{\mathsf{H}},j}(d_i)\,\mu\frac{v_{\mathsf{T},j}}{a}, & \text{if } L_i = \mathsf{H},\ l_i = \overline{\mathsf{H}},\ A(H_t, H_{t+1}, i); \\ \frac{\sigma_{\mathsf{H},j}(d_i)}{C(H_t,i)}(1 - \mu), & \text{if } L_i = \overline{\mathsf{H}},\ l_i = \mathsf{H},\ D_i = d_i,\ A(H_t, H_{t+1}, i); \\ 1 - \sigma_{\overline{\mathsf{H}},j}(d_i)\,\mu, & \text{if } L_i = l_i = \overline{\mathsf{H}},\ D_i = d_i,\ A(H_t, H_{t+1}, i); \\ 0, & \text{otherwise;} \end{cases} \tag{3}$$

where $C(H_t, i)$ is the number of queues with head-slot packets bound for output port d_i in state H_t, $\sigma_{\overline{\mathsf{H}},j}(d_i) = \sigma_{\overline{\mathsf{H}},j}$, and $\sigma_{\mathsf{H},j}(d_i) = \sigma_{\mathsf{H},j}$. A *space-conditional HOL factor* will be used in the computation of arrival rates. The stage-j space-conditional HOL factor, $f'_j(H_t, H_{t+1}, i)$, is given by (3) when $\sigma_{\overline{\mathsf{H}},j}(d_i) = \sigma_{\overline{\mathsf{H}},j}$, $d_i \neq 1$; $\sigma_{\overline{\mathsf{H}},j}(1) = 1$, and $\sigma_{\mathsf{H},j}(d_i) = \sigma_{\mathsf{H},j}$, $d_i \neq 1$; $\sigma_{\mathsf{H},j}(1) = 1$. The corresponding space-conditional transition probability is given by $T'_j(H_1, H_2) = \Pi_{i=1}^{a} f'_j(H_1, H_2, i)$.

3.3 QUEUE MODEL

Queue-model states are labeled $\mathsf{I}_{x,y}$ where $0 \leq x \leq m$, and $y \in \{\mathsf{T}, \overline{\mathsf{T}}, \phi\}$. The symbol x denotes the number of packets in the queue. The symbol $y = \mathsf{T}$ if the last occupied slot holds the

tail packet of a message, $y = \overline{\mathsf{T}}$ if the last occupied slot holds a non-tail packet of a message, and $y = \phi$ if $x = 0$. The probability that stage-j queue will be in state $\mathbf{l}_{x,y}$ is denoted $p_j(\mathbf{l}_{x,y})$.

Queue-model transition probabilities are a function of arrival rate, service rate, and expected message length. Define s_j to be the service rate, the probability that a stage-j HOL packet is able to move forward. Let $r_{i,j}$, $0 \leq i \leq m$, denote the arrival rate, the probability a new message will be ready to move into a stage-j queue given that the queue is in state $\mathbf{l}_{i,\mathsf{T}}$ or $\mathbf{l}_{0,\phi}$. Four distinct values of $r_{i,j}$ will be computed per queue: $r_{0,j}$ for an empty queue, $r_{m,j}$ for a full queue, $r_{m-1,j}$ for a queue with one slot free, and $r_{\mathsf{N},j}$ for a queue with 1 to $m-2$ packets. For notational simplicity, $r_{i,j}$, $0 < i < m-1$, will be used for $r_{\mathsf{N},j}$; this will be called the *normal* queue arrival rate.

The stationary probabilities must satisfy the following equations:

$$
\begin{aligned}
p_j(\mathbf{l}_{0,\phi}) &= p_j(\mathbf{l}_{0,\phi})(1-r_{0,j}) + p_j(\mathbf{l}_{1,\mathsf{T}})s_j(1-r_{1,j}) \\
p_j(\mathbf{l}_{1,\mathsf{T}}) &= p_j(\mathbf{l}_{0,\phi})r_{0,j}\mu + p_j(\mathbf{l}_{1,\mathsf{T}})[(1-s_j)(1-r_{1,j}) + s_j r_{1,j}\mu] + \\
&\quad p_j(\mathbf{l}_{1,\overline{\mathsf{T}}})s_j\mu + p_j(\mathbf{l}_{2,\mathsf{T}})s_j(1-r_{2,j}) \\
p_j(\mathbf{l}_{1,\overline{\mathsf{T}}}) &= p_j(\mathbf{l}_{0,\phi})r_{0,j}(1-\mu) + p_j(\mathbf{l}_{1,\mathsf{T}})s_j r_{1,j}(1-\mu) + p_j(\mathbf{l}_{1,\overline{\mathsf{T}}})s_j(1-\mu) \\
p_j(\mathbf{l}_{i,\mathsf{T}}) &= p_j(\mathbf{l}_{i-1,\mathsf{T}})(1-s_j)r_{i-1,j}\mu + p_j(\mathbf{l}_{i-1,\overline{\mathsf{T}}})(1-s_j)\mu + \\
&\quad p_j(\mathbf{l}_{i,\mathsf{T}})[(1-s_j)(1-r_{i,j}) + s_j r_{i,j}\mu] + \\
&\quad p_j(\mathbf{l}_{i,\overline{\mathsf{T}}})s_j\mu + p_j(\mathbf{l}_{i+1,\mathsf{T}})s_j(1-r_{i+1,j}) \\
p_j(\mathbf{l}_{i,\overline{\mathsf{T}}}) &= p_j(\mathbf{l}_{i-1,\mathsf{T}})(1-s_j)r_{i-1,j}(1-\mu) + p_j(\mathbf{l}_{i-1,\overline{\mathsf{T}}})(1-s_j)(1-\mu) + \\
&\quad p_j(\mathbf{l}_{i,\mathsf{T}})s_j r_{i,j}(1-\mu) + p_j(\mathbf{l}_{i,\overline{\mathsf{T}}})s_j(1-\mu) \\
p_j(\mathbf{l}_{m-1,\mathsf{T}}) &= p_j(\mathbf{l}_{m-2,\mathsf{T}})(1-s_j)r_{m-2,j}\mu + p_j(\mathbf{l}_{m-2,\overline{\mathsf{T}}})(1-s_j)\mu + \\
&\quad p_j(\mathbf{l}_{m-1,\mathsf{T}})[(1-s_j)(1-r_{m-1,j}) + s_j r_{m-1,j}\mu] + \\
&\quad p_j(\mathbf{l}_{m-1,\overline{\mathsf{T}}})s_j\mu + p_j(\mathbf{l}_{m,\mathsf{T}})s_j \\
p_j(\mathbf{l}_{m-1,\overline{\mathsf{T}}}) &= p_j(\mathbf{l}_{m-2,\mathsf{T}})(1-s_j)r_{m-2,j}(1-\mu) + p_j(\mathbf{l}_{m-2,\overline{\mathsf{T}}})(1-s_j)(1-\mu) + \\
&\quad p_j(\mathbf{l}_{m-1,\mathsf{T}})s_j r_{m-1,j}(1-\mu) + p_j(\mathbf{l}_{m-1,\overline{\mathsf{T}}})s_j(1-\mu) + p_j(\mathbf{l}_{m,\overline{\mathsf{T}}})s_j
\end{aligned}
$$

$$p_j(\mathbf{l}_{m,\mathsf{T}}) = p_j(\mathbf{l}_{m-1,\mathsf{T}})(1-s_j)r_{m-1,j}\mu + p_j(\mathbf{l}_{m-1,\overline{\mathsf{T}}})(1-s_j)\mu + p_j(\mathbf{l}_{m,\mathsf{T}})(1-s_j)$$

$$p_j(\mathbf{l}_{m,\overline{\mathsf{T}}}) = p_j(\mathbf{l}_{m-1,\mathsf{T}})(1-s_j)r_{m-1,j}(1-\mu) +$$

$$p_j(\mathbf{l}_{m-1,\overline{\mathsf{T}}})(1-s_j)(1-\mu) + p_j(\mathbf{l}_{m,\overline{\mathsf{T}}})(1-s_j)$$

for $2 \leq i \leq m-2$ and $1 \leq j \leq n$.

3.4 Computation of Rates

The queue- and HOL-arrival rates, service rate, and space probabilities are a function of HOL- and queue-model state probabilities.

The empty-queue arrival rate, $v_{\mathsf{E},j}$, $1 < j \leq n$, is computed from the stage-$(j-1)$ HOL-model state distribution. Let $\mathcal{H}_{\overline{1}}$ be the set of HOL states in which no HOL packets are destined for a particular port, without loss of generality, 1. The set is given by $\mathcal{H}_{\overline{1}} = \{\, H \mid H = \mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H},\ \forall_{1\leq i\leq a}\, d_i \neq 1 \,\}$. Similarly, let $\mathcal{H}_1$ be the set of states in which at least one HOL packet is destined for the port, $\mathcal{H}_1 = \{\, H \mid H = \mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H},\ \exists_{1\leq i\leq a}\, d_i = 1 \,\}$. An empty stage-$j$ queue at time t coincides with stage-$(j-1)$ HOL state $\mathcal{H}_{\overline{1}}$ at time $t-1$. The empty-queue arrival rate is then

$$v_{\mathsf{E},j} = \frac{\sum_{H_1 \in \mathcal{H}_{\overline{1}}} \sum_{H_2 \in \mathcal{H}_1} p_{j-1}(H_1) T_{j-1}(H_1, H_2)}{\sum_{H \in \mathcal{H}_{\overline{1}}} p_{j-1}(H)}$$

for $2 \leq j \leq n$. For the first stage $v_{\mathsf{E},1} = \lambda$, the arrival rate for new messages.

The quantity $v_{\mathsf{T},j}$ is computed so that the flow rate into a HOL slot is the network flow rate, ρ. This traffic is divided into three components. Flow entering: an empty HOL slot, a HOL slot containing a head packet, and a HOL slot containing a non-head packet. Let $\mathcal{H}_{\mathsf{E}}$ be the set of states in which a particular queue HOL slot, say 1, is empty, $\mathcal{H}_{\mathsf{E}} = \{\, H \mid H = \mathbf{H}_{((\phi,\mathsf{E}),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H} \,\}$. Let $\mathcal{H}_{\mathsf{H}'}$ be the set of states in which the HOL slot has a head packet that is not blocked by a message in progress, $\mathcal{H}_{\mathsf{H}'} = \{\, H \mid H = \mathbf{H}_{((d_1,\mathsf{H}),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H},\ \forall_{2\leq i\leq a}\, (d_i = d_1) \Rightarrow (l_i = \mathsf{H}) \,\}$. Let $\mathcal{H}_{\overline{\mathsf{H}}}$ be the set of states in which the HOL slot has a non-head

packet, $\mathcal{H}_{\overline{\mathsf{H}}} = \{\, H \mid H = \mathbf{H}_{((d_1,\overline{\mathsf{H}}),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H} \,\}$. Then $v_{\mathsf{T},j}$ is chosen so that

$$\sum_{H\in\mathcal{H}_{\mathsf{H}'}} p_j(H)\frac{\sigma_{\mathsf{H},j}}{C(H,1)}\left(\mu v_{\mathsf{T},j}+1-\mu\right)+\sum_{H\in\mathcal{H}_{\overline{\mathsf{H}}}} p_j(H)\sigma_{\overline{\mathsf{H}},j}\left(\mu v_{\mathsf{T},j}+1-\mu\right)+\sum_{H\in\mathcal{H}_{\mathsf{E}}} p_j(H)v_{\mathsf{E},j}=\rho$$

holds. Solving yields

$$v_{\mathsf{T},j}=\frac{\rho-\sum_{H\in\mathcal{H}_{\mathsf{H}'}} p_j(H)\frac{\sigma_{\mathsf{H},j}}{C(H,1)}(1-\mu)-\sum_{H\in\mathcal{H}_{\overline{\mathsf{H}}}} p_j(H)\sigma_{\overline{\mathsf{H}},j}(1-\mu)-\sum_{H\in\mathcal{H}_{\mathsf{E}}} p_j(H)v_{\mathsf{E},j}}{\sum_{H\in\mathcal{H}_{\mathsf{H}'}} p_j(H)\frac{\sigma_{\mathsf{H},j}}{C(H,1)}\mu+\sum_{H\in\mathcal{H}_{\overline{\mathsf{H}}}} p_j(H)\sigma_{\overline{\mathsf{H}},j}\mu}.$$

The stage-j queue-model service rate is equivalent to the stage-j HOL-model service rate. The HOL-model service rate is the probability that a HOL packet will advance. A non-head packet will advance if there is space; a head packet will advance if there is space and it wins contention. The service rate is given by

$$s_j=\frac{\sum_{H\in\mathcal{H}_{\mathsf{H}'}} p_j(H)\frac{\sigma_{\mathsf{H},j}}{C(H,1)}+\sum_{H\in\mathcal{H}_{\overline{\mathsf{H}}}} p_j(H)\sigma_{\overline{\mathsf{H}},j}}{\sum_{H\in(\mathcal{H}-\mathcal{H}_{\mathsf{E}})} p_j(H)}.$$

The empty-queue arrival probability is equivalent to the corresponding probability used in the HOL model, that is, $r_{0,j} = v_{\mathsf{E},j}$. The normal queue arrival probability, $r_{\mathsf{N},j}$, is found by considering the previous-stage HOL system. If a non-empty queue has less than $m-1$ items and the last occupied slot holds a tail packet then any packet that had entered the queue in the previous cycle was not blocked and had ended a message. This fact is used to obtain a previous-stage HOL-system distribution which is in turn used to find the arrival probability. Let $\mathcal{H}_{\mathsf{H}\to 1}$ be the set of states in which at least one HOL slot has a packet bound for a particular queue, say 1, and no HOL slot has a non-head packet bound for the queue. $\mathcal{H}_{\mathsf{H}\to 1} = \{\, H \mid H = \mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H},\ \exists_{1\le i\le a}\, d_i = 1,\ \forall_{1\le i\le a}\,(d_i=1)\Rightarrow(l_i=\mathsf{H}) \,\}$. Let $\mathcal{H}_{\overline{\mathsf{H}}\to 1}$ be the set of states in which no HOL slot has a non-head packet bound for the queue, $\mathcal{H}_{\overline{\mathsf{H}}\to 1} = \{\, H \mid H = \mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))} \in \mathcal{H},\ \forall_{1\le i\le a}\,(d_i=1)\Rightarrow(l_i=\mathsf{H}) \,\}$. Using the space-conditional transition probabilities,

$$r_{\mathsf{N},j}=\frac{\sum_{H_t\in\mathcal{H}}\sum_{H_{t+1}\in\mathcal{H}_{\mathsf{H}\to 1}} p_{j-1}(H_t)T'_{j-1}(H_t,H_{t+1})}{\sum_{H_t\in\mathcal{H}}\sum_{H_{t+1}\in\mathcal{H}_{\overline{\mathsf{H}}\to 1}} p_{j-1}(H_t)T'_{j-1}(H_t,H_{t+1})}$$

for $2 \leq j \leq n$. The arrival probability $r_{m-1,j}$ is computed so that the flow rate into the queue is ρ:

$$r_{m-1,j} = \frac{\rho - p_j(\mathbf{l}_{0,\phi})r_{0,j} - \sum_{x=1}^{m-2} p_j(\mathbf{l}_{x,\mathsf{T}})r_{x,j} - \sum_{x=1}^{m-1} p_j(\mathbf{l}_{x,\overline{\mathsf{T}}})}{p_j(\mathbf{l}_{m-1,\mathsf{T}})},$$

for $2 \leq j \leq n$. The full-queue arrival probability is found so that the fraction of time the previous-stage HOL model has a head packet bound for a particular queue matches the corresponding quantity in the queue model:

$$r_{m,j} = \frac{\sum_{H \in \mathcal{H}_{\mathsf{H}\to 1}} p_{j-1}(H) - p_j(\mathbf{l}_{0,\phi})r_{0,j} - \sum_{x=1}^{m-1} p_j(\mathbf{l}_{x,\mathsf{T}})r_{x,j}}{p_j(\mathbf{l}_{m,\mathsf{T}})},$$

for $2 \leq j \leq n$. For the first stage, $r_{i,1} = \lambda$ for $0 \leq i \leq m$.

Stage-j space probabilities $\sigma_{\mathsf{H},j}$ and $\sigma_{\overline{\mathsf{H}},j}$ are computed so the flow rate leaving a stage-j HOL slot is equal to the flow rate entering a stage-$(j+1)$ queue. The probability that there will be a head packet ready to enter a stage-$(j+1)$ queue is $p_{j+1}(\mathbf{l}_{0,\phi})v_{\mathsf{E},j+1} + \sum_{x=1}^{m} p_{j+1}(\mathbf{l}_{x,\mathsf{T}})r_{x,j+1}$; the probability that it is successful is

$$p_{j+1}(\mathbf{l}_{0,\phi})v_{\mathsf{E},j+1} + \sum_{x=1}^{m-1} p_{j+1}(\mathbf{l}_{x,\mathsf{T}})r_{x,j+1}.$$

This yields the head-packet space probability,

$$\sigma_{\mathsf{H},j} = 1 - \frac{p_{j+1}(\mathbf{l}_{m,\mathsf{T}})r_{m,j+1}}{p_{j+1}(\mathbf{l}_{0,\phi})v_{\mathsf{E},j+1} + \sum_{x=1}^{m} p_{j+1}(\mathbf{l}_{x,\mathsf{T}})r_{x,j+1}},$$

for $1 \leq j \leq n-1$. Similar reasoning is used for the non-head space probability,

$$\sigma_{\overline{\mathsf{H}},j} = 1 - \frac{p_{j+1}(\mathbf{l}_{m,\overline{\mathsf{T}}})}{\sum_{x=1}^{m} p_{j+1}(\mathbf{l}_{x,\overline{\mathsf{T}}})},$$

for $1 \leq j \leq n-1$. For the last stage, $\sigma_{\mathsf{H},n} = \sigma_{\overline{\mathsf{H}},n} = 1$.

The flow rate is determined by the arriving traffic and the fraction of time that traffic is not blocked:

$$\rho = \lambda \left(p_1(\mathbf{l}_{0,\phi}) + \sum_{i=1}^{m-1} p_1(\mathbf{l}_{i,\mathsf{T}}) \right) + \sum_{i=1}^{m-1} p_1(\mathbf{l}_{i,\overline{\mathsf{T}}}).$$

3.5 COMPUTATION OF DELAY

The *delay* of a message is defined to be the number of cycles that the head packet is in the network. The *normalized delay* is defined as the delay divided by the number of stages. The *waiting time* of a message in a queue is defined to be the number of cycles that the head packet spends in the queue. (This includes what others call service time.) The delay then is the sum of the waiting times. The total time a message spends in the network is the delay plus the message length minus one.

The expected waiting time is found by first computing, for $0 \leq i < m$, the waiting time, $w_{i,j}$, for a head packet arriving at a stage-j queue that had i packets in the previous cycle. The expected waiting time is the weighted sum over i. Let

$$s'_j = \frac{\sum_{H \in \mathcal{H}_{\mathsf{H}'}} p_j(H)\sigma_{\mathsf{H},j}/C(H,1)}{\sum_{H \in (\mathcal{H}-\mathcal{H}_{\mathsf{E}}-\mathcal{H}_{\overline{\mathsf{H}}})} p_j(H)}$$

be the service probability of a HOL packet that is a head packet. The expected waiting time of a head packet at a stage-j queue HOL slot is then $1/s'_j$. The expected waiting time of a HOL packet of unknown type is $1/s_j$. If the queue had i packets in the cycle before a message arrived then $w_{i,j} = i/s_j + 1/s'_j - 1$. The expected waiting time, w_j, is then given by

$$w_j = \frac{p_j(\mathbf{l}_{0,\phi})\frac{r_{0,j}}{s'_j} + \sum_{i=1}^{m-1} p_j(\mathbf{l}_{i,\mathsf{T}})r_{i,j}(i/s_j + 1/s'_j - 1)}{p_j(\mathbf{l}_{0,\phi})r_{0,j} + \sum_{i=1}^{m-1} p_j(\mathbf{l}_{i,\mathsf{T}})r_{i,j}}.$$

The normalized delay is then $\sum_{j=1}^{n} w_j/n$.

3.6 ANALYSIS PROCEDURE

Stationary distributions for the HOL and queue models are obtained through iteration, using the following procedure. State probabilities are initialized uniformly. That is, if a state model has x states then the probability of each state is initialized to $1/x$. Space, arrival, and service probabilities are initialized to 0.5. Other initializations are possible; all those tested yielded the same results. After the second iteration new values for the state probabilities, service and arrival

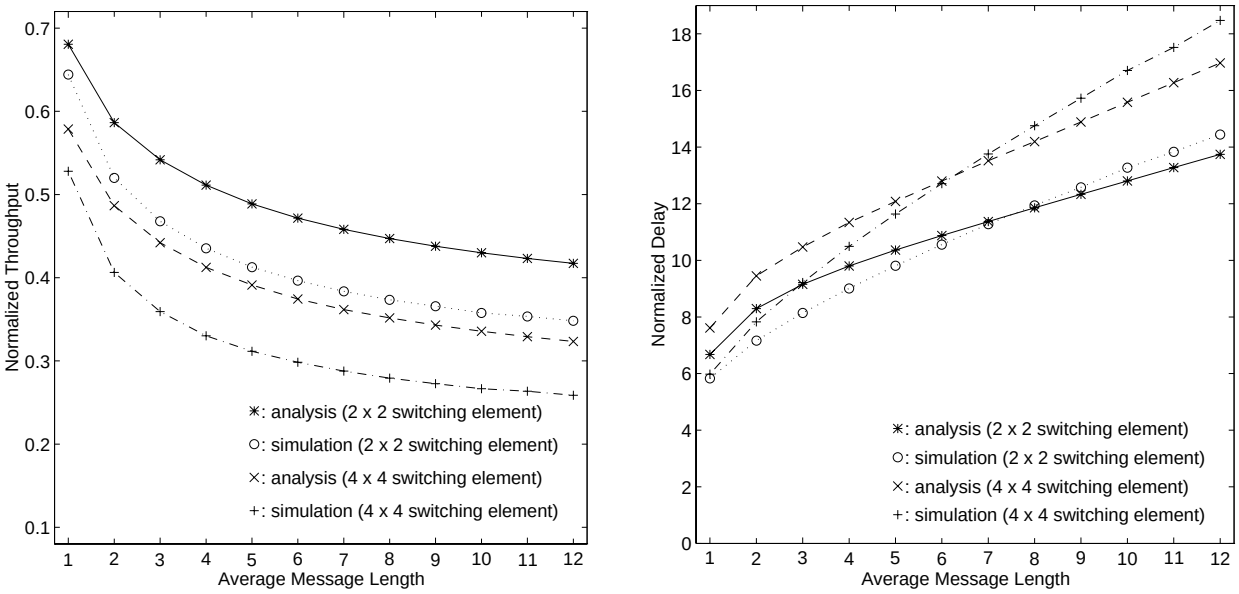

Figure 2. Throughput and delay v. message length for $(5, 2, 8)$ and $(5, 4, 8)$ networks, $\lambda = 1$.

rates, and space probabilities are computed using the average of values computed in the previous two iterations. (At the first two iterations initial values are used.) Iteration proceeds until the difference between values computed for quantities in consecutive iterations is sufficiently small.

4 RESULTS

The analysis was tested by comparing its predictions against those of a simulator. Comparisons of predicted throughput and delay were made for a variety of network and traffic models. Network size, arrival rate, queue size, switching-element size, and message size were varied.

The simulator uses the same network and traffic model as the analysis. Simulations were performed for 100,000 cycles; simulator output includes delay and throughput. Confidence intervals (95%) were computed assuming that simulator throughput and delay computed for a series of identical runs are normally distributed. The confidence intervals are extremely small. The analysis was performed for less than 1000 iterations in most cases. The number of iterations was chosen so that corresponding probabilities differed by less than 10^{-6} in the last two iterations of

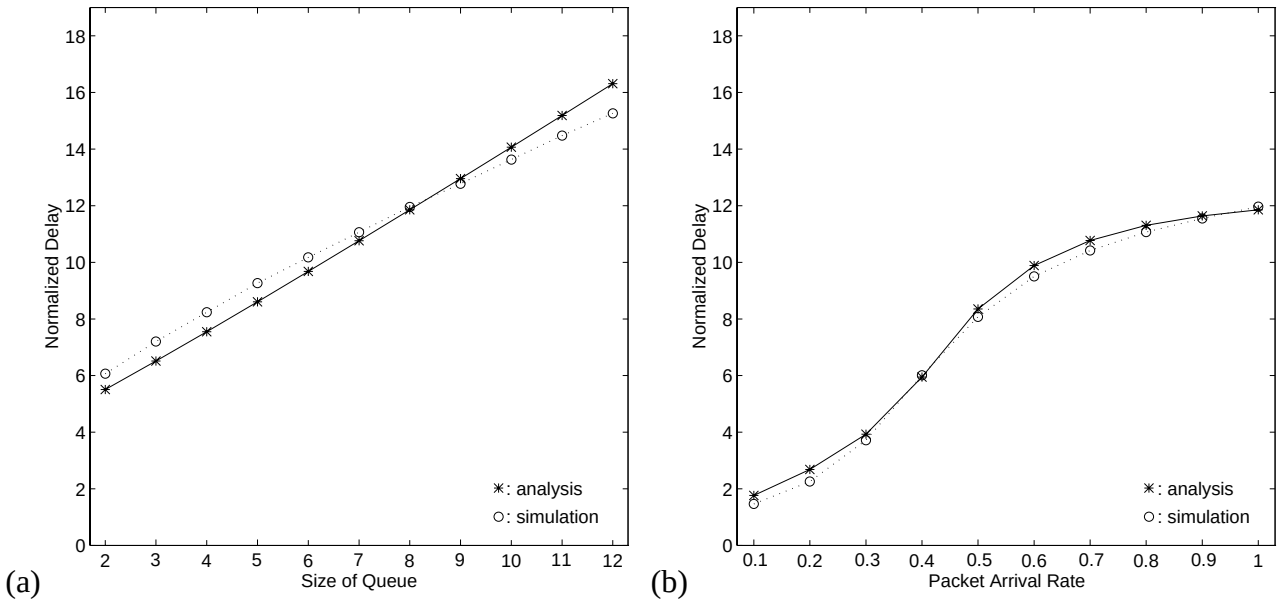

Figure 3. (a) Delay v. arrival rate in $(5, 2, 8)$ networks. (b) Delay v. queue size in $(5, 2, m)$ networks for $\lambda = 1$. Both for message length 8.

each analysis.

The prediction of message-length effects on networks using 2×2 and 4×4 switching elements can be seen in Figure 2. The figure shows the normalized throughput and delay for networks offered saturating traffic, that is, $\lambda = 1$. The effect of message size is clearly modeled. As with other analyses of this type, the throughput is overestimated. The delay in simulated and analyzed systems closely match.

The prediction of delay at varying arrival rates for an average message length of 8 are plotted in Figure 3(*a)*. The delay computed closely matches simulations. At higher arrival rates the throughput computed (not shown) is too high. As can be seen in Figure 3(*b)*the effect of queue size on network performance is predicted. In that figure, delay is plotted against queue size for $(5, 2, m)$ networks. Note that the queue size ranges from smaller-than to larger-than the average message length. Delay is accurately predicted for smaller queue sizes, but diverges for larger queues.

The effect of network size was also tested. As with unit-message-length analyses, the throughput prediction is increasingly overestimated as the number of stages increases. The delay prediction, in contrast, remains close to the delays obtained from simulation.

5 CONCLUSIONS

A geometrically distributed-message-length banyan-network analysis has been presented. This is one of the few banyan network analyses that consider anything other than fixed-length messages. This is of value because in real parallel computers and communication networks message sizes vary. In the analysis, the banyan-network switching elements are captured by two state models: one modeling a single queue, the other modeling all of a switching-element's queue heads. Banyan networks with finite queue sizes and arbitrary switching-element size can be analyzed.

The analysis was tested against simulations. The results show that message-length effects are effectively modeled. In particular, the negative impact that long messages have on network performance is predicted.

6 APPENDIX

Let π_ι and π_o be permutations of switching-element input and output labels, respectively. Then states $H_\alpha = \mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}$ and H_β are said to be *equivalent* if there exists an input-label permutation π_ι and an output-label permutation π_o such that $\Pi(H_\alpha) = H_\beta$ where

$$\Pi(\mathbf{H}_{((d_1,l_1),(d_2,l_2),\ldots,(d_a,l_a))}) = \mathbf{H}_{((\pi_\mathrm{o}(d_{\pi_\iota(1)}),l_{\pi_\iota(1)}),(\pi_\mathrm{o}(d_{\pi_\iota(2)}),l_{\pi_\iota(2)}),\ldots,(\pi_\mathrm{o}(d_{\pi_\iota(a)}),l_{\pi_\iota(a)}))}$$

and $\pi(x)$ is the symbol to which x is mapped under permutation π.

Lemma: For all $1 \leq j \leq n$ and $H_1, H_2 \in \mathcal{H}$ transition probability $T_j(H_1, H_2) = T_j(\Pi(H_1), \Pi(H_2))$ where Π is any of the switching-element mappings described above.

Inspection of equations (1-3) will reveal that permuting switching-element labels will have no effect. For example, consider predicate $D_i = d_i$ in (3). Clearly $D_i = d_i \iff \pi_\mathrm{o}(d_{\pi_\iota(j)}) = \pi_\mathrm{o}(D_{\pi_\iota(j)})$, where $\pi_\iota(j) = i$. Other references to switching-element inputs and outputs are also independent of absolute or relative position and so are unaffected by the permutations. A detailed proof is omitted.

7 REFERENCES

[1] L. R. Goke and G. J. Lipovski, "Banyan networks for partitioning multiprocessor systems," in *Proceedings of the International Symposium on Computer Architecture,* 1973, pp. 21–28.

[2] J.Y. Hui, "Switching and traffic theory for integrated broadband networks," Boston: Kluwer Academic Publishers, 1990.

[3] Y. C. Jenq, "Performance analysis of a packet switch based on single–buffered banyan network," *IEEE Journal on Selected Areas in Communications,* vol. 1, no. 6, pp. 1014–1021, June 1983.

[4] C. P. Kruskal and M. Snir, "A unified theory of interconnection network structure," *Theoretical Computer Science,* vol. 48, pp. 75–94, 1986.

[5] C. P. Kruskal, M. Snir, and A. Weiss, "The distribution of waiting times in clocked multistage interconnection networks," *IEEE Transactions on Computers,* vol. 37, no. 11, pp. 1337–1352, November 1988.

[6] F. T. Leighton, "Introduction to parallel algorithms and architectures: arrays * trees * hypercubes," Palo Alto: Morgan Kaufmann, 1992.

[7] Y. Mun and H. Y. Youn, "Performance analysis of finite buffered multistage interconnection networks," *IEEE Transactions on Computers,* vol. 43, no. 2, pp. 153-162, February 1994.

[8] J. A. Patel, "Performance of processor-memory interconnections for multiprocessors," *IEEE Transactions on Computers,* vol. 30, pp. 771–780, 1981.

[9] H. Yoon, K. Y. Lee, and M. T. Liu, "Performance analysis of multibuffered packet–switching networks in multiprocessor systems," *IEEE Transactions on Computers,* vol. 39, no. 3, pp. 319–327, March 1990.

List of Figures

Loose Ends

TABLE OF CONTENTS