On the Average Delay for Routing Subject to Independent Deflections

Bruce Hajek, *Fellow, IEEE*, and Rene L. Cruz, *Senior Member, IEEE*

***Abstract*—Consider a packet walking along a directed graph with each node having two edges directed out. The packet is headed towards one of N destinations, chosen according to a probability distribution p. At each step, the packet is forced to use a nonpreferred edge with some probability q, independently of past events. Using information theory and sequential analysis, it is shown that the mean number of steps required by the packet to reach the destination is, roughly, at least $H(p)/(1-h(q))$, where h is the binary entropy function and $H(p)$ is the entropy (base two) of p. This lower bound is shown to be asymptotically achievable in the case the packet always begins at a fixed node. Also considered is the maximum, over all pairs of nodes in a graph, of the mean transit time from one node to the other. The work is motivated by the search for graphs which work well in conjunction with deflection routing in communication networks.**

***Index Terms*—Routing, interconnection networks, deflection routing, packet switching.**

I. Introduction

CONSIDER the directed graph shown in Fig. 1. The edges of the graph represent unidirectional communication lines and the nodes of the graph represent packet switches. Suppose a packet is initially placed at node s_o, and that it is destined for some node θ. A preferred edge out of each node is specified as a function of θ, as shown for example in Fig. 1. The time axis is divided into equal length slots. Suppose the packet is at some node v at the beginning of a time slot. If $v \neq \theta$ then the packet attempts to traverse during the slot the preferred edge leading from v. The attempt is successful with probability $1-q$, independently of past events, where q is a given constant. If the attempt is not successful, the packet is deflected, meaning that it traverses the nonpreferred edge out of node v.

The objective of this paper is to consider graphs such that the average time it takes a packet to reach a randomly specified destination, in the face of deflections as just described, is small. Section II provides a lower bound on the mean time required for arbitrary graphs. Section III shows that the lower bound can be asymptotically met if either the destinations are sets of nodes or if arbitrarily large node in-degrees are permitted. Section IV allows any pair of nodes to serve as the source and destination of the packet. It provides examples of graphs for which the maximum, over all pairs of nodes, of the expected transit time between the nodes is reasonably small. In the remainder of this section, the specific model and notation will be introduced, and motivation for the model will be given.

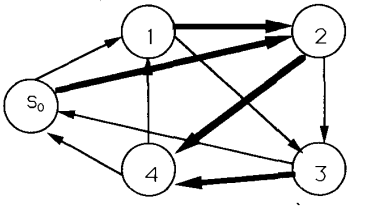

Fig. 1. Graph $G = (V, E)$ with an initial node s_o indicated. Preferred edges for destination node $\theta = 4$ are indicated in bold.

Let $G = (V, E)$ be a directed graph with set of nodes V and set of edges E. An edge is permitted to loop from a node back to the same node. The *in-degree* (respectively, *out-degree*) of a node is the number of edges leading into (out of) the node. Suppose each node v has out-degree two, and let $V(v)$ denote the set of nodes at the head ends of the edges leading out of v. For notational simplicity assume that the two edges leading from a node lead to distinct nodes, so that for each v, $V(v)$ contains two nodes. Assume that $N \geq 1$ and that $(D(\theta) : 1 \leq \theta \leq N)$ is a collection of disjoint, nonempty subsets of V, and that $(p(\theta) : 1 \leq \theta \leq N)$ is a probability distribution. Let s_o denote a fixed node in V.

Suppose that a packet initially placed at s_o moves along a path in the graph, one node at a time. A random variable Θ with distribution $(p(\theta) : 1 \leq \theta \leq N)$ is observed at time zero, and the goal is for the packet to visit a node in $D(\Theta)$ as soon as possible. Let $X(k, \theta)$ denote the node visited at time k given that $\Theta = \theta$, so that $X(0, \theta) = s_o$ and $X(k+1, \theta) \in V(X(k, \theta))$. The probability distribution of $X = (X(k, \theta) : k \geq 0)$ is governed by a routing function $R = (R_\theta(v) : 1 \leq \theta \leq N, v \in V)$ and a deflection probability q, where $R_\theta(v) \in V(v)$ and $0 \leq q < 0.5$. Given $X(0, \theta)$, $X(1, \theta)$, $\cdots$, $X(k, \theta)$ with $X(k, \theta) = v$, the preferred next node is $R_\theta(v)$. However, assume that the actual next node is $\overline{R}_\theta(v)$ with probability q, where $\overline{R}_\theta(v)$ is the node in $V(v)$ different from $R_\theta(v)$. To make this more explicit, let $B = (B(k) : k \geq 1)$ be a sequence of independent random variables with $P[B(k) = 1] = q = 1 - P[B(k) = 0]$ which is also independent of Θ. Define $X = (X(k, \theta) : k \geq 0)$ by

Manuscript received June 1, 1990; revised December 1991. This work was supported by the National Science Foundation under Contracts NSF ECS 90 04355 and NSF NCR 89 04029. This work was presented in part at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 24–28, 1991.

B. Hajek is with the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering, University of Illinois, 1101 W. Springfield, Urbana, IL 61801.

R. L. Cruz is with the Department of Electrical and Computer Engineering, University of California, San Diego, Mail Code R-007, LaJolla, CA 92093.

IEEE Log Number 9203008.

$X(0,\theta) = s_o$ and for $k \geq 1$,

$$X(k,\theta) = \begin{cases} R_\theta(X(k-1,\theta)), & \text{if } B(k) = 0, \\ \overline{R}_\theta(X(k-1,\theta)), & \text{if } B(k) = 1. \end{cases} \quad (1)$$

Thus, $\{B_k = 1\}$ indicates that the kth move of the packet is a deflection.

Let, for each θ with $1 \leq \theta \leq N, T_\theta$ be a stopping time for X. Thus, for any θ, T_θ is a non negative integer-valued function of X such that for any k, the event $\{T_\theta \leq k\}$ is determined by $(X(j,\theta) : 0 \leq j \leq k)$. In view of (1), T_θ is also a stopping time for B. A reasonable choice for T_θ is

$$T_\theta = \min\{M, \min\{k : X(k,\theta) \in D(\theta)\}\},$$

for some constant M. Another choice, which is appropriate for networks in which a node cannot be both a destination node and a transit node, is

$$T_\theta = \min\{k : X(k,\theta) \in D^*\}, \quad (2)$$

where $D^* = D(1) \cup D(2) \cup \cdots \cup D(N)$. Our aim in Sections II and III is to study the problem of finding (G, s_o, R, D, T) so that $E[T_\Theta]$ is small while $P[X(T_\Theta, \Theta) \in D(\Theta)]$ is close to one. In Section IV, the important variation of this problem is considered in which any pair of nodes in the graph can serve as the packet source and destination, and the maximum, over all such pairs, of the mean transit time is considered.

The problems considered in this paper were formulated in order to help find graphs that work well with *deflection routing*. Deflection routing, originally called *hot-potato routing* [1], is a technique for maintaining bounded buffers in a packet-switched communication network. If, due to congestion at a switch, not all packets can be sent out along shortest paths to their destinations, some packets are sent out on other links. The penalty is an increase in the distance traveled by packets, and the reward is the simplicity of switch design resulting from the absence of large buffers and routing tables. Traditional store-and-forward networks use extensive computation at the nodes to determine packet routes in order to use transmission bandwidth sparingly. In contrast, deflection routing leads to simple switches by making liberal use of transmission bandwidth. Since the penalty for deflection is probably severe if long propagation delays are involved, deflection routing is often primarily considered for networks with a small physical diameter, such as those in a multiprocessor computer system or a packet switch for telecommunications. This paper concentrates on networks with 2×2 switches, consistent with a recent experimental demonstration of optical 2×2 switches supporting deflection routing [2]. A destination set $D(\theta)$ may contain more than one node if it corresponds to a set of nodes connected to a trunk group in a telecommunications switch [3].

Several authors have given methods to approximately analyze networks with deflection routing. For example, see [4]–[8]. The methods typically lead to a random walk model similar to the one considered here, where the deflection probability q is a function of the congestion in the network. Roughly speaking, q quantifies the impact of deflections by other packets in the network. The experience of many authors is that the approximate analysis matches simulations quite well when the traffic is balanced in some sense and certain independence assumptions are nearly true. This paper starts with the approximate model and thereby skips the analysis of multiple packet interactions. The hope is that progress on the problem formulated here (especially the variation described in Section IV) will suggest new network designs that are effective for deflection routing. The graphs described in Section IV are a step in that direction. Novel designs that are promising according to the random walk model can later be tested under multiple packet interaction by simulation and the existing approximate analysis methods.

II. Lower Bound on Delay

Let $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ for $0 \leq x \leq 1$ and let $H(a)$ denote the binary entropy of a probability distribution a, $H(a) = -\sum_{\theta=1}^{N} a(\theta)\log_2 a(\theta)$. Given $(\beta_\theta : 1 \leq \theta \leq N)$ with $0 \leq \beta_\theta \leq 1$, define β and a probability distribution $\hat{p}$ by

$$\beta = \sum_{\theta=1}^{N} p(\theta)\beta_\theta \quad \text{and} \quad \hat{p}(\theta) = \frac{p(\theta)(1-\beta_\theta)}{1-\beta}.$$

Theorem 1: Suppose that $H(\hat{p}) \geq 1.45$ and that

$$P[X(T_\theta, \theta) \notin D(\theta)] \leq \beta_\theta, \quad (3)$$

for $1 \leq \theta \leq N$. Then,

$$E[T_\Theta] \geq \frac{(1-\beta)(H(\hat{p}) - \log_2 H(\hat{p}) + \log_2(1-h(q)) - 1.45) - h(\beta)}{1-h(q)}. \quad (4)$$

If, in addition, T_θ is given by (2) for each θ, then

$$E[T_\Theta] \geq \frac{(1-\beta)H(\hat{p}) - h(\beta)}{1-h(q)}. \quad (5)$$

Remarks: a) If β_θ is equal to β for all θ, the $p \equiv \hat{p}$. If, in addition, $p(\theta) \equiv 1/N$, then $H(\hat{p}) = \log_2 N$. b) In the limit as $\beta \to 0$ and $H(\hat{p}) \to +\infty$ with q fixed, the right-hand side of inequality (4) is asymptotically equivalent to $H(\hat{p})/(1-h(q))$.

More notation and a lemma will be given before the theorem is proved. Let $\mathcal{P}$ denote the set of paths of finite length in G which start at s_o. Each path p thus has the form $p = (s_o = v_o, v_1, \cdots, v_k)$ where $v_i \in V$, and the length of such a path, denoted $l(p)$, is k. The case $k = 0$ is permitted. Note that there are 2^k paths of length k in $\mathcal{P}$ for $k \geq 0$ since each node in G has out degree 2.

Define $\mathcal{B}$ by

$$\mathcal{B} = \{\partial\} \cup \{(b_1, \cdots, b_k) : k \geq 1, b_i \in \{0,1\}\},$$

where ∂ is a symbol that, informally, denotes a binary sequence of length zero. The length, $l(b)$, of a sequence $b = (b_1, \cdots, b_k)$ in $\mathcal{B}$ is k.

A path $p = (x_o, x_1, \cdots, x_k)$ of length k is said to end in $D(\theta)$ if $x_k \in D(\theta)$. For $1 \leq \theta \leq N$, let

$$\tilde{D}(\theta) = \{p \in \mathcal{P} : p \text{ ends in } D(\theta)\}.$$

Given θ with $1 \le \theta \le N$, there in a one-to-one correspondence between $\mathcal{B}$ and $\mathcal{P}$ suggested by (1). Specifically, given $b = (b_1, \cdots, b_k)$ in $\mathcal{B}$, define $\gamma_\theta(b)$ to be the path $p = (x_o, x_1, \cdots, x_k)$ where $x_o = s_o$ and, for $1 \le j \le k$,

$$x_j = \begin{cases} R_\theta(x_{j-1}), & \text{if } b_j = 0, \\ \overline{R}_\theta(x_{j-1}), & \text{if } b_j = 1. \end{cases}$$

Also let $\gamma_\theta(\partial)$ equal (s_o), the path in $\mathcal{P}$ of length zero. Then γ_θ is a bijection from $\mathcal{B}$ to $\mathcal{P}$.

Equation (3) can be rewritten as

$$P[(X(0,\theta), \cdots, X(T_\theta, \theta)) \notin \tilde{D}(\theta)] \le \beta_\theta.$$

which is equivalent to

$$P[(B(1), \cdots, B(T_\theta)) \notin \gamma_\theta^{-1}(\tilde{D}(\theta))] \le \beta_\theta. \tag{6}$$

Recall that under probability measure P, the variables $B(k)$, $k \ge 1$, are independent with $P[B(k) = 1] = q = 1 - P[B(k) = 0]$. Thus, under P the sequence B takes values in $\{0,1\}^{\mathcal{Z}+}$ where $\mathcal{Z}+$ denotes the set of positive integers. Trivially, B also takes values in the larger set $\{0,1,2\}^{\mathcal{Z}+}$, and the new stopping time given by $\min\{T_\theta, \min\{k : B_k = 2\}\}$ is naturally defined for any B in that set. Since this new stopping time is an extension of T_θ, the notation T_θ will, henceforth, be used to denote it.

Let $\delta \ge 0$ and P_o be the probability measure defined on the Borel subsets of $\{0,1,2\}^{\mathcal{Z}+}$ so that under measure P_o, the coordinate variables $(B(k) : k \ge 1)$ are independent and, for each k,

$$P_o[B(k) = i] = \begin{cases} 2^{-1-\delta}, & \text{if } i = 0 \quad \text{or } i = 1, \\ 1 - 2^{-\delta}, & \text{if } i = 2. \end{cases}$$

Lemma 1: Let $1 \le \theta \le N$ and set

$$\alpha_\theta = P_o[(B(1), \cdots, B(T_\theta)) \in \gamma_\theta^{-1}(\tilde{D}(\theta))]. \tag{7}$$

Then,

$$E[T_\theta] \ge \frac{1}{1 - h(q) + \delta}\left\{-h(\beta_\theta) + (1 - \beta_\theta)\log_2\frac{1}{\alpha_\theta}\right\}. \tag{8}$$

Proof: Since the right-hand side of inequality (8) is otherwise negative, assume, without loss of generality, that $\alpha_\theta + \beta_\theta < 1$. Equations (6) and (7) show that the event

$$\{(B(1), \cdots, B(T_\theta)) \in \gamma_\theta^{-1}(\tilde{D}(\theta))\}$$

corresponds to a sequential test for the hypothesis H : "B is governed by distribution P" versus the hypothesis H_0 : "B is governed by distribution P_o." Equation (6) states that the probability of deciding H_o is true given H is true is at most β_θ and (7) states that the probability of deciding H is true given H_o is true is α_θ. To apply Wald's elementary theory of sequential hypothesis testing, note also that

$$E\left[\log_2\frac{f(B(1))}{f_o(B(1))}\right] = 1 - h(q) + \delta,$$

where f and f_o, respectively, are the probability mass functions for $B(1)$ under measures P and P_o. A basic theorem of sequential analysis [9, Theorem 2.39][1] therefore yields that

$$E[T_\theta] \ge \frac{1}{1 - h(q) + \delta} \cdot \left\{(1 - \beta_\theta)\log_2\frac{1 - \beta_\theta}{\alpha_\theta} + \beta_\theta \log_2\frac{\beta_\theta}{1 - \alpha_\theta}\right\}. \tag{9}$$

The nonnegative term, $-\beta_\theta \log_2(1 - \alpha_\theta)$, on the right-hand side of (9) can be dropped to obtain (8), and the lemma is proved. □

Proof of Theorem 1: Using the lemma and the definition of $\hat{p}$, and applying Jensen's inequality for the concave function h, yields

$$\begin{aligned} E[T_\Theta] &= \sum_\theta p_\theta E[T_\theta] \\ &\ge \frac{1}{1 - h(q) + \delta}\sum_\theta p(\theta)\left[-h(\beta_\theta) + (1 - \beta_\theta)\log_2\frac{1}{\alpha_\theta}\right] \\ &\ge \frac{1}{1 - h(q) + \delta}\left\{-h(\beta) + (1 - \beta)\sum_\theta \hat{p}(\theta)\log_2\frac{1}{\alpha_\theta}\right\}. \end{aligned} \tag{10}$$

Next a lower bound for the quantity $\sum_\theta \hat{p}(\theta)\log_2 1/\alpha_\theta$ is given. Since the mapping γ_θ is a bijection and preserves lengths,

$$\begin{aligned} \alpha_\theta &= P[(B(1), \cdots, B(T_\theta)) \in \gamma_\theta^{-1}(\tilde{D}(\theta))] \\ &\le \sum_{b \in \gamma_\theta^{-1}(\tilde{D}(\theta))} 2^{-(1+\delta)l(b)} = \sum_{p \in \tilde{D}(\theta)} 2^{-(1+\delta)l(p)}. \end{aligned} \tag{11}$$

The sets $\tilde{D}(\theta), 1 \le \theta \le N$, are disjoint subsets of $\mathcal{P}$, and there are exactly 2^k paths in $\mathcal{P}$ of length k. Thus,

$$\sum_{\theta=1}^{N} \alpha_\theta \le \sum_{p \in P} 2^{-(1+\delta)l(p)} = \sum_{k=0}^{\infty} 2^{-k\delta} = \frac{1}{(1 - 2^{-\delta})}. \tag{12}$$

Viewing (12) as a constraint on $(\alpha_\theta : 1 \le \theta \le N)$ under which $\sum_\theta \hat{p}(\theta)\log_2 1/\alpha_\theta$ is to be minimized yields

$$\sum_\theta \hat{p}(\theta)\log_2\frac{1}{\alpha_\theta} \ge H(\hat{p}) + \log_2(1 - 2^{-\delta}). \tag{13}$$

Applying (13) in (10) (and writing H for $H(\hat{p})$),

$$E[T_\Theta] \ge \frac{(1 - \beta)[H + \log_2(1 - 2^{-\delta})] - h(\beta)}{1 - h(q) + \delta}. \tag{14}$$

Choose $\delta > 0$ so that $1 - 2^{-\delta} = (1 - h(q))/H$, which is always possible since it is assumed that $H \ge 1.45$. Substituting

[1] Since use of Theorem 2.39 in Siegmund's book requires that equality hold in (6), appeal to the fact that the right-hand side of (9) is decreasing in β_θ for $\alpha_\theta + \beta_\theta < 1$.

for δ in the right side of (14), bounding the denominator:

$$\begin{aligned} 1 - h(q) + \delta &= 1 - h(q) - \log_2\left(1 - \frac{1-h(q)}{H}\right) \\ &= (1-h(q))\left(1 + (\log_e 2)^{-1} \cdot \sum_{k=1}^{\infty} \frac{(1-h(q))^{k-1}}{kH^k}\right) \\ &\le (1-h(q))\left(\sum_{k=0}^{\infty} \frac{1}{(H\log_e 2)^k}\right) \\ &= (1-h(q))\left(\frac{1}{1-\frac{1}{H\log_e 2}}\right), \end{aligned}$$

and using the fact $1/(\log_e 2) < 1.45$ yields inequality (4).

Turning to the proof of Equation (5), assume that (2) holds for all θ. The proof is similar to that of (4), but now take $\mathcal{P}$ to be the set of all finite length paths in G of the form $(s_0 = v_0, v_1, \cdots, v_k)$ where $v_i \notin D^*$ if $i < k$ and $v_k \in D^*$, and let $\delta = 0$. Since, with this new definition, no path in $\mathcal{P}$ is an initial part of any other path in $\mathcal{P}$, Kraft's inequality implies the following replacement for (12):

$$\sum_{\theta=1}^{N} \alpha_\theta \le \sum_{p \in P} 2^{-l(p)} \le 1. \tag{15}$$

Equation (13) can be strengthened to

$$\sum_{\theta} \hat{p}(\theta) \log_2 \frac{1}{\alpha_\theta} \ge H(\hat{p}). \tag{16}$$

Applying Equation (16) in (10) yields (5), and the theorem is proved. □

Remark: Inequality (4) and its proof are more complicated than (5) and its proof since, in the later case, the set $\mathcal{P}$ has a prefix property required for the Kraft inequality. The technique used in the former case is close to one of the methods introduced by Leung–Yan–Cheong and Cover [10] for variable length source coding without a prefix condition on the code.

III. A Converse

The following theorem shows that the lower bound on $E[T_\Theta]$ provided in Theorem 1 is asymptotically the best possible as $N \to \infty$ with $p(\theta) \equiv 1/N$, q fixed, and β small.

Theorem 2: Given $0 \le q < 0.5$ and $\beta > 0$, for N sufficiently large there exists (G, s_o, D, R, T) with $P[X(T_\theta, \theta) \notin D(\theta)] \le \beta$ and $T_\theta \le ((1+\beta)\log_2 N)/(1-h(q))$ for $1 \le \theta \le N$.

Proof: Let q with $0 \le q < 1/2$ be fixed, and consider the memoryless binary symmetric channel with cross-over probability q. The Shannon capacity of the channel is $1-h(q)$ bits per channel use. Thus, given $\beta > 0$, if N is sufficiently large there exists a block code with N codewords of length L, where $L = \lfloor((1+\beta)\log_2 N)/(1-h(q))\rfloor$, such that the error probability is less than β for each codeword. Explicitly, the code can be denoted by $(c_\theta : 1 \le \theta \le N)$, where $c_\theta = (c_\theta(1), \cdots, c_\theta(L)) \in \{0,1\}^L$, and there are corresponding decoding sets $(D(\theta) : 1 \le \theta \le N)$, which are disjoint subsets of $\{0,1\}^L$. The requirement on the error probability is that, for $1 \le \theta \le N$,

$$P[(c_\theta(1) \oplus B(1), \cdots, c_\theta(L) \oplus B(L)) \notin D(\theta)] \le \beta,$$

where $a \oplus b$ is the modulo 2 sum of a and b.

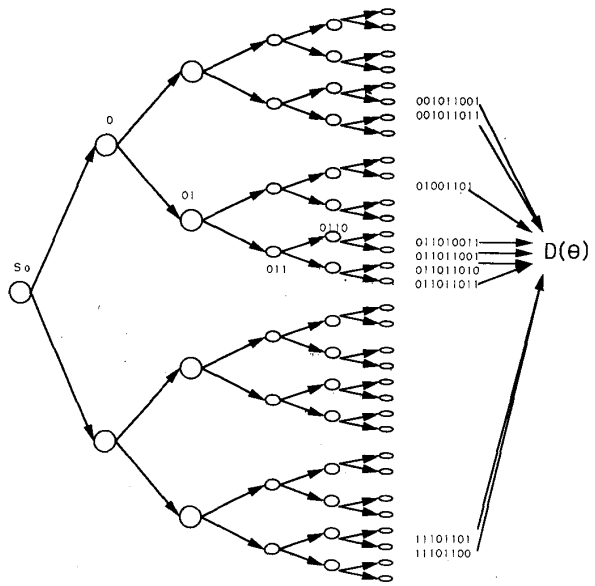

Fig. 2. Tree graph of the type satisfying bound in Theorem 2. Destination set $D(\theta)$ indicated might correspond to the codeword $c_\theta = 011011011$.

A system (G, s_o, D, R, T) is determined by the code as follows. The graph is a full directed binary tree of depth L (see Fig. 2). The root node is the source node s_o and the destination sets $(D(\theta) : 1 \le \theta \le N)$ are subsets of the set of leaves. Each of the 2^k nodes that are k steps from s_o is indexed by a binary sequence $b_1 b_2 \cdots b_k \in \{0,1\}^k$, and for $1 \le k \le L-1$, $V(b_1 \cdots b_k) = \{b_1 b_2 \cdots b_k 0, b_1 b_2 \cdots b_k 1\}$. In addition, $V(s_o) = \{0,1\}$. Choose $T_\theta \equiv L$ for all θ, and let the decoding sets $(D(\theta) : 1 \le \theta \le N)$ be identical to the destination sets, which are subsets of nodes, with each node L steps from s_o. The routing function for a given destination θ is determined by codeword c_θ. Specifically, the choice of edge for the kth step of the packet is given by $c_\theta(k)$, the kth bit of $R_\theta(b_1 \cdots b_k) = b_1 \cdots b_k c_\theta(k)$. The description of the system (G, s_o, D, R, T) is now complete, and by design $T_\theta = L \le ((1+\beta)\log_2 N)/(1-h(q))$ for $1 \le \theta \le N$.

The node reached by the packet at time T_θ is given by

$$x(T_\theta, \theta) = (c_\theta(1) \oplus B(1), \cdots, c_\theta(L) \oplus B(L)),$$

so that $X(\theta, T_\theta) \notin D(\theta)$, if and only if the codeword c_θ is incorrectly received for the error sequence $(B(1), \cdots, B(L))$. Thus, by the choice of the code, $P[X(T_\theta, \theta) \notin D(\theta)] \le \beta$ for $1 \le \theta \le N$. □

Remark: A slight modification of the system constructed in the proof of Theorem 2 yields a graph in which there is precisely one node in each destination set. It is obtained by concentrating each destination set down to a single node in one step, thereby only increasing the delay by one. Of course, the destination nodes must have large in-degree.

IV. Constructions with Any Node Eligible to be a Source or Destination

So far we have considered graphs in which the packet always starts at the same node and the destinations are possibly sets of nodes. In this, section, we suppose the packet can start at any node of the graph and be destined for any other node in the graph. Of interest is $\max E[T]$, the maximum, over all source-destination pairs, of the mean transit time. Since maximizing produces a larger value than fixing the source and averaging over destinations, Theorem 1 (with $\beta = 0$ and $p(\theta) = 1/N$) yields that

$$\max E[T] \geq \frac{\log_2 N - \log_2 \log_2 N + \log_2 (1 - h(q)) - 1.45}{1 - h(q)}. \tag{17}$$

While we are as yet unable to construct graphs that asymptotically meet this lower bound on $\max E[T]$, the relatively simple graphs described below come fairly close.

As a starting point, consider a standard shuffle-exchange graph with parameter n, defined as follows. The nodes are labeled by binary sequences of length n, and the two edges leading out of node $b_n b_{n-1} \cdots b_1$ lead to the nodes $b_{n-1} \cdots b_1 0$ and $b_{n-1} \cdots b_1 1$. That is, the label of the node at the head of an edge is obtained from the label of the node at the tail of the edge by dropping the lead bit and appending a new bit. In the absence of deflections, a packet starting at any node can reach any specified destination node by traversing at most n edges. Examining the sequence of labels of the nodes traversed, we see that the label of the destination node is "shifted in from the right."

If along the way the packet is deflected, a wrong bit is shifted in. Often this causes the packet to lose most of the progress it has made, because the new distance-to-go will typically be close to n. Since the probability of being not deflected in n consecutive steps is $(1-q)^n$, the mean time for a packet to reach its destination grows at least as fast as $(1-q)^{-n}$, which is far from linear in n. We shall discuss three ways to obtain smaller mean delays.

A. Shuffle-Exchange Graph with Greenberg/Goodman Modification

Greenberg and Goodman [5] suggested a way to modify an arbitrary graph $G = (V, E)$ with all nodes having in-degree and out-degree equal to two. For $u, v \in V$ let $\mathrm{dist}(u, v)$ denote the minimum number of steps needed to get from u to v (in the absence of deflections). Assume that $\mathrm{dist}(u, v)$ is finite for all u, v (i.e., the graph is *strongly connected)* and let $\mathrm{diam}(G) = \max_{u,v \in V} \mathrm{dist}(u, v)$.

The corresponding modified graph is $\hat{G} = (\hat{V}, \hat{E})$. All nodes in it also have in-degree and out-degree 2, $|\hat{V}| = 4|V|$, and $\mathrm{diam}(\hat{G}) \leq 2\mathrm{diam}(G) + 2$. Each node in G has two corresponding nodes in $\hat{G}$, and each *edge* in G has a corresponding node in $\hat{G}$, as indicated in Fig. 3. Formally, $\hat{V} = (V \times \{0, 1\}) \cup E$, and if $e = (u, v) \in E$, then $(u, 0)$, $(u, 1)$, $(v, 0)$, $(v, 1)$ and e are all in $\hat{V}$, and $((u, 0), e), (e, (v, 0)), ((v, 1), e)$, and $(e, (u, 1))$ are edges in $\hat{G}$. This specifies $\hat{G}$. Obviously $|\hat{V}| = 4|V|$. The fact $\mathrm{diam}(\hat{G}) \leq 2\mathrm{diam}(G) + 2$ is not hard to establish, starting with the observation that $\mathrm{dist}((u, 0), (v, 0)) \leq 2\mathrm{dist}(v, u)$ and $\mathrm{dist}((u, 1), (v, 1)) \leq 2\mathrm{dist}(v, u)$.

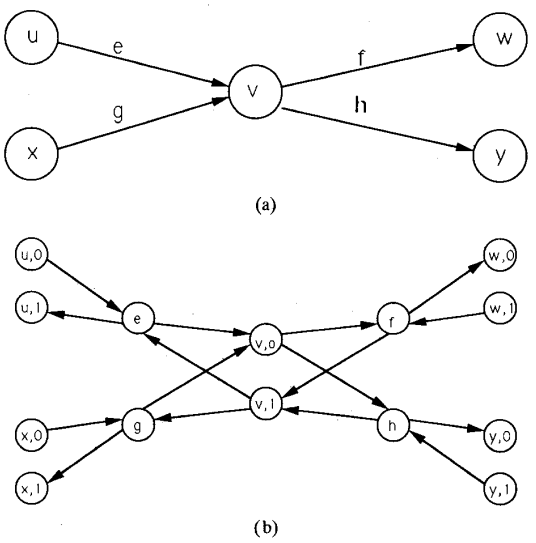

Fig. 3. (a) Fragment of a directed graph G with nodes all having in-degree and out-degree 2. (b) Fragment of the corresponding graph $\hat{G}$.

Each edge in $\hat{G}$ is contained in a cycle of length four. Thus, if a packet is deflected, then in three additional steps it can return to the point of deflection [5]. In fact, each edge in $\hat{G}$ is contained in *two* distinct cycles of length four, so that in one of the three additional steps either one of the two out-going edges can be selected. Call that step a *don't care* step.

Let a, $b \in \hat{V}$ and consider a packet starting at node a, destined for node b, and subject to independent deflections with probability q at each step. Assume that *stubborn* routing is used, in which the packet tries to follow a particular shortest length path from a to b. Whenever the packet is deflected, it insists on returning to the point of deflection before continuing. If it gets deflected multiple times, the packet always tries to return to the point of the most recent deflection that it has not yet revisited. Thus, under stubborn routing, the path the packet ultimately follows is nominally a shortest length path, which can be augmented by four-step excursions which, in turn, can be augmented by more such excursions, and so on.[2]

A *true deflection* is a deflection that does not happen during a don't care step. Clearly $T = \mathrm{dist}(a, b) + 4D$, where T is the total number of steps to reach b(so $\beta = 0$), $\mathrm{dist}(a, b)$ is

[2]Stubborn routing does not quite fit the model defined in the introduction since in stubborn routing the preferred outgoing edge at a node does not depend only on the node. Instead, the preferred outgoing edge depends on the path that was used in arriving at the node. However, a standard argument of dynamic programming [11, pp. 71–78] shows that there is another routing strategy that does fit the original model for which $E[T]$ is less than or equal to what it is for stubborn routing.

computed relative to graph $\hat{G}$, and D is the total number of true deflections during the packet's journey to b. Consider the first k steps, and let D_k denote the number of them that were true deflections and let A_k denote the number of them that were not don't cares. Then $E[D_k] = qE[A_k]$. (In fact, $(D_k - qA_k)_{k\geq 0}$ is a martingale.)

Initially there are $\text{dist}(a,b)$ opportunities for true deflections, and each true deflection creates three more opportunities for true deflections. Thus $A_k \leq \text{dist}(a,b) + 3D_k$ so that $E[D_k] = qE[A_k] \leq q\text{dist}(a,b) + 3qE[D_k]$. Assuming now that $q < 1/3$, $E[D_k] \leq q\text{dist}(a,b)/(1-3q)$. By the monotone convergence theorem, $E[D] = \lim_{k\to\infty} E[D_k]$ so that $E[D] \leq q\text{dist}(a,b)/(1-3q)$. Therefore,

$$E[T] \leq \text{dist}(a,b) + 4E[D] \leq \text{dist}(a,b)\left(1 + \frac{4q}{1-3q}\right) \leq \frac{\text{diam}(\hat{G})(1+q)}{1-3q}. \quad (18)$$

Let us now take G to be the shuffle-exchange graph with parameter n defined at the beginning of the section. Then $|V| = 2^n$ and $\text{diam}(G) = n$ so that $|\hat{V}| = 2^{n+2}$ and $\text{diam}(\hat{G}) \leq 2n+2$. Equation (18) applied to $\hat{G}$ yields that (let $N = |\hat{V}|$)

$$\max E[T] \leq \frac{2+2q}{1-3q}\log_2 N, \quad (19)$$

for $0 \leq q < 1/3$.

B. A Graph with Backspace Edges

When q tends to zero the upper bound on $\max E[T]$ given in (19) tends to $2\log_2 N$, which is twice the worst case transit time for a shuffle-exchange graph. The graph described next offers average transit time nearer to $\log_2 N$ for small values of q. It has parameters k and l, which are both positive integers, and it can be arranged in three stages, as pictured in Fig. 4.

The first stage contains l columns of M nodes each, where $M = 2^{kl}$, and the second and third stage each contain one column of $2M$ nodes. The nodes within a column in stage 1 are labeled by binary sequences of length $n = kl$. The edges out of each column in the first stage of the network form a shuffle interconnection with two edges leading out of each node. More precisely, the two edges out of node $b_n b_{n-1}\cdots b_1$ in column j of stage 1 lead to the nodes $b_{n-1}b_{n-2}\cdots b_1 0$ and $b_{n-1}b_{n-2}\cdots b_1 1$ in column $j+1$ of stage 1, where $1 \leq j \leq l-1$. The nodes in stages 2 and 3 are indexed by binary sequences of length $n+1$. The two edges out of node $b_n b_{n-1}\cdots b_1$ in column l of stage 1 lead to the nodes $b_{n-1}b_{n-2}\cdots b_1 0 b_n$ and $b_{n-1}b_{n-2}\cdots b_1 1 b_n$ in stage 2. Node $b_n b_{n-1}\cdots b_1 b_0$ in stage 2 has an edge leading to node $b_n b_{n-1}\cdots b_1 b_0$ in stage 3 and an edge leading to node $b_l b_{l-1}\cdots b_1 b_n b_{n-1}\cdots b_{l+1} b_0$ in stage 3 (called a backspace edge and corresponding to an l-fold right circular shift on the n most significant bits). Finally, node $b_n b_{n-1}\cdots b_1 b_0$ in stage 3 has an edge leading to node $b_n b_{n-1}\cdots b_1 b_0$ in stage 2 (such edge is the preferred edge out of the node in backspace mode only) and an edge leading to node $b_n b_{n-1}\cdots b_1$ in column 1 of stage 1.

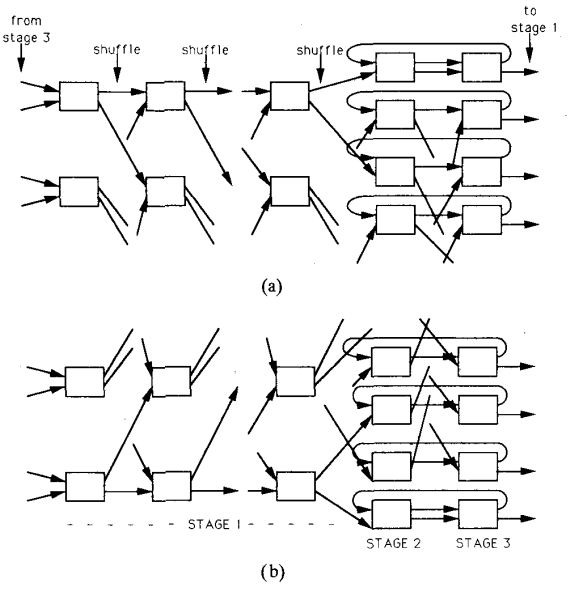

Fig. 4. Graph with backspace edges.

In the absence of deflections, the packet repeatedly cycles through stages 1, 2, and 3 in order without traversing backspace edges or the edges going from stage 3 directly back to stage 2. The destination label is shifted in from the right, one bit per column, when the packet passes through stage 1. In the absence of deflections, the packet can reach its destination before completing $k+1$ cycles. However, if the packet suffers one or more deflections in traversing stage 1, it later uses a backspace edge between stages 2 and 3, corresponding to shifting out the l bits shifted in stage 1 (including the erroneous ones). Deflections can occur at any node of the graph, even when the packet attempts to traverse a backspace edge, and we assume that stubborn routing is used.

It can be shown that

$$l - 1 + \left(\log_2 \frac{N}{l+4}\right) D(q,l) \leq \max E[T] \leq \left(l + \log_2 \frac{N}{l+4}\right) D(q,l), \quad (20)$$

where N is the number of nodes in the network. The constant $D(q,l)$ is finite, if and only if there exists a vector $(x_1, x_2, x_3, y_2, y_3)$ with positive coordinates satisfying

$$\begin{aligned} x_1 &= l + \left(1 - (1-q)^l\right)(y_2 + x_3 + x_1), \\ x_2 &= 1 + q(x_3 + x_1 + x_2), \\ x_3 &= 1 + q(x_2 + x_3), \\ y_2 &= 1 + q(y_3 + y_2), \\ y_3 &= 1 + q(x_1 + y_2 + y_3), \end{aligned} \quad (21)$$

and when such a vector exists, $D(q,l) = (x_1 + x_2 + x_3)/l$. Here, x_i $(i = 1, 2, \text{or } 3)$ denotes the mean number of steps needed to correctly pass through stage i in normal mode, and y_i $(i = 2 \text{ or } 3)$ denotes the mean number of steps to correctly pass through stage i in backspace mode. These five means

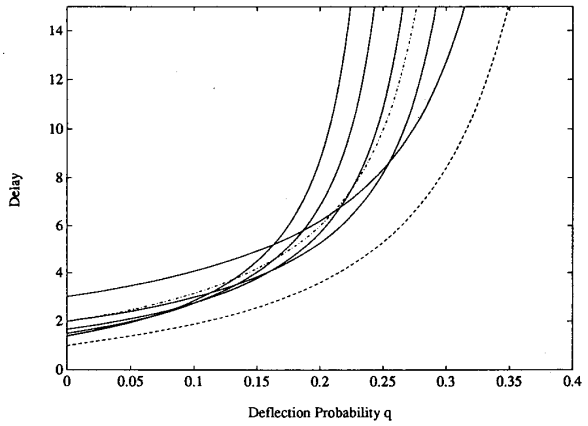

Fig. 5. Asymptotic values of $\max E[T]/\log_2 N$ versus q, for the shuffle-exchange graph with the Greenberg/Goodman modification (dot–dash), and for the graphs with error correcting stages for $1 \leq l \leq 5$ (solid curves, value $(l+2)/l$ at $q = 0$). For comparison, the lower bound $1/(1-h(q))$ is also indicated (dashed).

include contributions due to all possible excursions caused by deflections. Equations (21) were obtained by conditioning on whether at least one deflection occurs, and can be solved numerically by straight-forward numerical iteration, or algebraically.

Thus, $\max E[T]/\log_2 N$ converges to $D(q,l)$ as N (equivalently k) tends to infinity with q and l fixed. Some values of $D(q,l)$ are shown in Fig. 5. Note that the value of l that minimizes $D(q,l)$ increases as q decreases, reflecting the fact that the opportunity to take a backspace edge need be offered less frequently when q is small.

C. Reduction of Node Deflection Probability

The constructions in the previous two subsections yield $\max E[T]$ of size $O(\log N)$ for q fixed, with $q < 1/3$ in the first instance and with $l = 1$ and $q < 0.3966$ in the second instance. We show in this subsection how to extend this result to any fixed q with $q < 1/2$. The basic idea is indicated in Fig. 6. It shows a single 2-output node constructed by interconnecting three 2-output nodes. Trite calculations show that if the three component nodes deflect packets independently with probability q, then under the obvious routing strategy the composite node deflects packets with probability q_{new} given by $q_{\text{new}} = (2q^2 - q^3)/(1 - q + q^2)$. It is easy to check that $q_{\text{new}} < q$ for $0 < q < 1/2$. In addition, the average delay $D(q)$ suffered by a packet in passing though the composite node satisfies

$$D(q) = \frac{2 + q(1-q)}{1 - q(1-q)} < 3. \tag{22}$$

Thus, if each node in a graph is replaced by such a composite node and if the destination is the first component node of a composite node, then the mean delay is at most a factor three larger than the mean delay for the original graph with q replaced by the smaller q_{new}. For example, by replacing each node of the graph constructed in the first subsection by a composite node, we arrive at a graph such that the mean delay from any source to the first component node of any composite node is at most $(D(q)(2 + 2q_{\text{new}}))/(1 - 3q_{\text{new}}) \log_2 N$, where N is the number of nodes in the new graph. It can be shown that the mean delay for this graph is at most larger by a constant (depending on q but not N) if *any* node is eligible to be the destination node. This procedure may be iterated, with all the nodes again being replaced by three interconnected nodes each. The result, after a fixed number of iterations (depending on q) are finite functions $C_1(q)$ and $C_2(q)$ and corresponding graphs, such that $\max E[T] < C_1(q) \log N + C_2(q)$ for $0 \leq q < 1/2$.

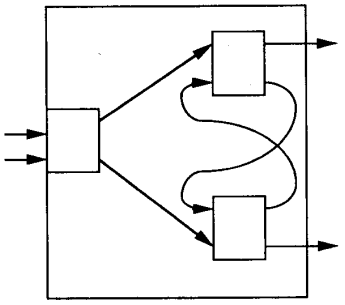

Fig. 6. Composite node constructed from three component nodes.

V. Discussion

We leave open the problem of closing the gap between the normalized asymptotic maximum delay for graphs under independent deflections, and the lower bound, as indicated in Fig. 5. Theorem 2 shows that the lower bound cannot be improved unless somehow the constraint that any node is possibly a source node is incorporated into the derivation.

On a more practical note, return to the case discussed in the introduction in which deflections are caused by interactions among multiple packets. The graph $\hat{G}$ derived from a shuffle-exchange graph by the Greenberg/Goodman modification apparently has the following property: If many packets are simultaneously routed over the graph using deflection routing, and if the sources and destinations are independent and uniformly distributed, then the traffic will tend to be uniformly distributed over the edges of the graph. In particular, the independence assumptions that have been used in the analysis of deflection routing will likely be fairly accurate. On the other hand, the graphs with an error correcting stage have two distinguished columns of nodes in stages 2 and 3. While assuming independent deflections at these nodes for real traffic may not be accurate, the fact that there are twice as many nodes per column in the last two stages should alleviate congestion there.

Acknowledgment

The authors wish to thank A. Barron, J. Brassil, A. Greenberg, and J. Wolf for useful suggestions.

References

[1] P. Baran, "On distributed communication networks," *IEEE Trans. Commun. Syst.*, vol. 12, pp. 1–9, 1964.

[2] D. J. Blumenthal, K. Y. Chen, J. Ma, R. J. Feurstein, and J. R. Sauer, "A deflection routing 2×2 photonic switch for computer interconnections," submitted to *Electron. Lett.,* Oct. 1991.
[3] R. L. Cruz, "The statistical data fork: A class of broadband multichannel switches," *IEEE Trans. Commun.*, vol. 40, pp. 1625–1634, Oct. 1992.
[4] J. Brassil and R. L. Cruz, "Nonuniform traffic in the Manhattan street network," in *Proc. 1991 IEEE ICC,* pp. 1647–1651, June 1991.
[5] A. G. Greenberg and J. Goodman, "Sharp approximate models of adaptive routing in mesh networks," in *Teletraffic Analysis and Computer Performance Evaluation,* O. Boxma, J. W. Cohen, and H. C. Tijms, Eds. Amsterdam: Elsevier, 1986, pp. 255–270. (Revision to appear in *IEEE Trans. Commun.*)
[6] A. G. Greenberg and B. Hajek, "Approximate analysis of deflection routing in hypercube networks," *IEEE Trans. Commun.,* vol. 40. pp. 1070–1081, June 1992.
[7] A. Krishna and B. Hajek, "Performance of shuffle-like switching networks with deflection," in *Proc. IEEE INFOCOM 90, Conf. on Comput. Commun.,* IEEE Computer Society Press, June 1990, pp. 473–480.
[8] D. H. Lawrie and D. A. Padua, "Analysis of message switching with shuffle-exchanges in multiprocessors," in *The Proceedings of the Workshop on Interconnection Networks for Parallel and Distributed Processing,* 1980, pp. 116–123. Reprinted in *Interconnection Networks,* Wu and Feng, Eds. New York: IEEE Computer Society Press, 1984.
[9] D. Siegmund, *Sequential Analysis.* New York: Springer-Verlag, 1985.
[10] S. K. Leung-Yan-Cheong and T. M. Cover, "Some equivalences between Shannon entropy and Kolmogorov complexity," *IEEE Trans. Inform. Theory,* vol. IT-24, pp. 331–338, May 1978.
[11] P. R. Kumar and P. Varaiya, *Stochastic Systems.* Englewood Cliffs, NJ: Prentice-Hall, 1986.