Deflection Routing in Hypercube Networks

Albert G. Greenberg, *Member, IEEE,* and Bruce Hajek, *Fellow, IEEE*

***Abstract*—An approximate analysis of the transient and steady state behavior of deflection routing in hypercube networks is presented, under a uniform traffic model. In deflection routing congestion causes packets admitted to the network to be temporarily misrouted rather than buffered or dropped. Our approximations show that deflection routing performs remarkably well in hypercube networks, for small as well as large networks and for the whole range from light to heavy load. Simulations suggest that the approximations are quite accurate.**

I. Introduction

QUEUEING theory deals with jobs contending for resources. In many important problems each resource can be granted to at most one job at a time. Typically, it is assumed that jobs not granted their desired resource are either 1) queued, or 2) rejected and cleared. However, if there are as many resources as jobs then there is a third option: Jobs not granted their desired resource may be granted some other resource temporarily. We say that such jobs are *deflected.*

This third option makes sense in certain communication networks where jobs correspond to data packets (messages) and resources to communication lines between nodes in the network. In *deflection routing* [1], [2], [6], [9], [10], [12] nodes attempt to route each packet along a shortest path to its destination. However, when a node holds two or more packets whose desired paths call for the same communication line the node grants the line to one of the packets and grants other lines to the other packets. In this way, congestion causes packets admitted to the network to be misrouted temporarily, in contrast with traditional schemes where such packets might be buffered or dropped.

In this paper we consider deflection routing in *hypercube* networks, under a stochastic model where the destinations of the packets generated at each node are distributed uniformly among the other nodes in the network. In a hypercube, there are $N = 2^d$ nodes for some $d \geq 1$ and there is a communication line (edge) between nodes i and j $(0 \leq i, j \leq N-1)$ if the binary representations of i and j are the same except at exactly one of the d bit positions. Hypercube networks have long been proposed for the communications backbone for parallel processors, where a small cluster of processors is associated with each node and processors in different clusters communicate by routing packets through the network. Several such parallel processors have been built [5], some of which rely on a type of deflection routing.[1]

In Section II we describe the network, the routing algorithm, and the stochastic model governing the generation and handling of new packets. We then consider the transient and equilibrium behavior of the network. In our model new packets may be blocked, that is, not admitted to the network, but we show in Section IV how networks with input queueing rather than blocking can also be analyzed. Statistics of interest include the probability that a typical packet is blocked (possibly as a function of time of arrival), the average delay of an accepted packet, and the distribution of how far a packet involved in a typical deflection is from its destination at the time of the deflection. The network can be modeled by a single finite-state Markov chain, but the number of states increases very rapidly with the network dimension. In Section III we give an approximate analysis and associated asymptotics, which judging from simulation results are quite accurate. Syzmanski [13], in a paper appearing well after this paper was submitted, gives a similar approximate analysis. His equations are at once more general and more complicated. He does not study asymptotics.

Deflection routing is essentially the same as hot potato routing as originally defined by Baran [1, pp. 6–7]. We use the terminology "deflection routing" instead of "hot potato routing" only because the later terminology has come to have a much broader meaning. Other early descriptions of deflection routing (with no special terminology) include [9] and [12]. Borodin and Hopcroft [2] posed an interesting problem of deflection routing in hypercube networks where each node initially holds one packet, and the packet destinations form a permutation of the N node indices. No new packets are ever generated. The problem is to route the N packets to their destinations, and the statistic of interest is the longest packet delay. In [2] it was reported that simulation results suggest that, for every permutation, a version of deflection routing achieves a longest delay of $c \log N$ for some small constant $c > 0$. Maxemchuk [10]/[11] proposed deflection routing in Manhattan street networks (two-dimensional meshes where neighboring rows or columns alternate in direction). Greenberg and Goodman [3] gave an approximate analysis of deflection routing in these networks, under a uniform traffic model similar to that described here. Tan et al. [14] studied a type of deflection routing in shuffle exchange and related networks via simulation. Additional work

Paper approved by the Editor for Wide Area Networks of the IEEE Communications Society. Manuscript received August 25, 1989; revised December 27, 1990 and August 6, 1991.

The work of B. Hajek was supported by the National Science Foundation under contract NSF ECS 83 52030 with matching funds provided by AT&T. This paper was presented at the TIMS Summer Meeting, Osaka, Japan, June 1989.

A. G. Greenberg is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

B. Hajek is with the Coordinated Science Laboratory, University of Illnois, Urbana-Champaign, IL 61801.

IEEE Log Number 9108051.

[1] The connection machine of thinking machines Inc. [7] and the HEP parallel processor [12] of Denelcor Inc.

on deflection routing is summarized in the bibliography of [8].

II. Models

We consider a hypercube network with 2^d nodes, $d \geq 1$. All deflection routing schemes require a rule for resolving the conflicts that arise when a subset of packets at a node contend for a smaller set of links. We introduce *one-pass* deflection routing, which embodies a very simple probabilistic conflict resolution rule.

A. Network Operation

The nodes operate synchronously: the time axis is divided into slots and each link can relay one packet per time slot. To describe the network operation, consider the operation of a node during a time slot. Any packets that were destined to the node and reached it in a previous slot are assumed to be removed before the beginning of the slot. At the beginning of a slot there are

- U *continuing* packets (received from other nodes during the previous slot and destined for other nodes) at the node, and
- V *new* packets (generated locally) offered to the node.

Since at most one packet arrived per incoming link, $U \leq d$. Of the $U+V$ packets, $(U+V-d)^+$ are blocked and dropped from consideration and $V \wedge (d-U)$ are accepted where $a \wedge b$ denotes the minimum of a and b. The blocked packets are all new ones—continuing packets are never blocked. Hence, $(U+V) \wedge d$ packets are to be transmitted by the node during the slot.

Under *one-pass* deflection routing, these packets are assigned to outgoing edges in the following way. The packets are considered one at a time, in random order, with all orders being equally likely. When a given packet is considered, the node examines which outgoing links are on shortest paths to the packet's destination (call these *preferred links*-there are i of them if the packet is i hops from its destination) and sees if any of those are free, i.e., not already assigned to another packet. If at least one preferred link is free, the packet is assigned to a preferred, free link at random, all choices being equally likely. Otherwise, the packet is assigned to a free (but not preferred) link, all choices being equally likely, and we say the packet is *deflected*.

Our model of network operation does not include synchronization errors or propagation delays. The one pass rule was chosen for its simplicity and its freedom from deadlock (where packets repeatedly deflect in such a way that some never reach their destination). While this rule is not the best possible, i) our simulations, analytic approximations, and asymptotics show that it will be hard to do significantly better, and ii) the one-pass rule is not particularly difficult to implement.

B. Offered Traffic

The following model for offered traffic will be assumed. Initially, the network holds no packets. Of course a node can transmit at most d packets per slot. Let $0 \leq v \leq d$ and suppose that the number of new packets offered to a node during a slot has the binomial distribution $B(d, v/d)$ where the notation $B(n,p)$ represents the binomial distribution with parameters $n \geq 1$ and $0 \leq p \leq 1$. Assume that the number of packets offered is independent from node to node and slot to slot. Also, assume that the destination of a new packet arriving at any given node is uniformly distributed over the set of $2^d - 1$ nodes obtained by excluding the node the packet arrived on. The destinations of all packets are chosen independently. Thus,

$$q(i) = \binom{d}{i} \Big/ \left(2^d - 1\right); \qquad 1 \leq i \leq d \tag{2.1}$$

gives the distribution and $d/\left(2\left(1-2^{-d}\right)\right)$ the mean of the initial distance of a packet from its destination. Finally, assume that the new packets that are blocked in a given slot are rejected and cleared.

Thus, the parameters of the network and traffic models are d and v. We remark that while v can be as large as d, the maximum long term throughput of the network is less than 2. That is so because accepted packets must traverse at least $d/2$ links on the average and there are d unidirectional links in the network per node. Thus, fewer than 2 packets per node can be accepted in the long run. Our asymptotic analysis of approximate state equations suggests that, even with the added burden due to deflections, throughput near 2 packets per node per slot can be sustained—see Theorem 3.1.

III. Analysis

We wish to determine both the transient and equilibrium behavior of the network. Statistics of interest include the probability that a typical packet is blocked (possibly as a function of time of offering), the average delay of an accepted packet, and the distribution of how far a packet involved in a typical collision is from its destination at the time of the deflection.

The arrivals of packets on different incoming links of a node are not necessarily independent, though we expect that under our uniform traffic model they should be nearly so. We will derive an approximate performance analysis by pretending that arrivals on different incoming links of a node during any slot are independent. Specifically we assume, for any node and time slot:

Approximation 3.1: On any given one of the node's d incoming links a continuing packet is or is not received independently of whether packets are received on the other links.

Approximation 3.2: Consider a single packet (new and accepted, or continuing) present at the node at this slot, and suppose that the packet is i hops from its destination where $1 \leq i \leq d$, so that it has i preferred links. At the moment the packet is to be assigned an outgoing link, the indentities of these i links are randomly, uniformly distributed, independently of the links already assigned to other packets.

Using these approximations, first we will find the acceptance and deflection probabilities at one node during one time slot. The calculations will then be used to derive an approximate analysis of the whole network, at any given time t (Section III-B) and in equilibrium (Section III-C). Next, we will investigate

the behavior of the routing algorithm for large hypercube networks (Section III-D). We will see that the formulae simplify significantly in the limit as the hypercube dimension $d \to \infty$. The asymptotics indicate that the equilibrium performance of the routing algorithm becomes optimal in this limit.

A. One Node, One Time Slot Calculation

Consider a single node and its behavior in one time slot. Consider any incoming link and let

$$m = \Pr[\text{a continuing packet is received on the link during this slot .}]$$

In following subsections m is replaced by a time dependent variable m_t to be determined by induction, or by a stationary value $\bar{m}$ to be determined by a fixed point equation. By Approximation 3.1, the number U of continuing packets present has the binomial distribution $B(d, m)$. The number, V, of new packets offered at the node has the binomial distribution $B(d, v/d)$, and is independent of U. Since the number of accepted new packets is $(d - U) \wedge V$, the proba bility that a typical new packet is accepted is given by

$$a(m, d, v) = \frac{E[(d - U) \wedge V]}{v}. \tag{3.1}$$

Note that $a(m, d, v)$ can be readily numerically computed for given (m, d, v) by averaging over the $(d+1)^2$ possible values of (U, V).

Next we will find $p(i, m, d, v)$ (resp., $p_o(i, m, d, v)$) which is defined to be the probability that a typical *continuing* packet (resp., typical accepted new packet) is deflected, given that the node is i hops away from the packet's destination. It will be seen that both $p(i, m, d, v)$ and $p_o(i, m, d, v)$ are readily numerically computed for given (i, m, d, v).

Consider a typical continuing packet and suppose that it is i hops from its destination. The number of other continuing packets U_o has the binomial distribution $B(d-1, m)$ and so the packet must compete with $(U_o + V) \wedge (d-1)$ other packets. Given (U_o, V), the number R of other packets assigned before the one considered is equally likely to be any of the $(U_o + V + 1) \wedge d$ values in $\{0, 1, \cdots, (U_o + V) \wedge (d-1)\}$. The typical packet is blocked if and only if the i links it prefers are all in the set of R links already assigned to other packets. By Approximation 3.2, the probability the packet is deflected given $R = j$ is $\left(\frac{j}{d}\right)\left(\frac{j-1}{d-1}\right)\cdots\left(\frac{j-i+1}{d-i+1}\right)$, or equivalently, $\frac{j_i}{d_i}$ where $n_i = n(n-1)\cdots(n-i+1)$. Thus,

$$p(i, m, d, v) = E[H((U_o + V) \wedge (d-1), d, i)] \tag{3.2}$$

where

$$H(k, d, i) = \frac{1}{(k+1)} \sum_{j=1}^{k} \frac{j_i}{d_i}. \tag{3.3}$$

Now consider a typical new offered packet that is i hops from its destination. The number of other offered new packets V_o has the binomial distribution $B(d-1, v/d)$. We will show that

$$p_o(i, m, d, v) = \frac{1}{a(m, d, v)} E\Big[\frac{(1 + V_o) \wedge (d - U)}{1 + V_o} \cdot H((U + V_o) \wedge (d-1), d, i)\Big]. \tag{3.4}$$

First, given (U, V_o), the first term inside the expectation on the right-hand side is the conditional probability that the packet is accepted, and the second term is the conditional probability that it is deflected given that it is accepted [the reasoning is similar to that for (3.2)]. Thus, the expectation is the probability the packet is accepted and deflected. This establishes the equation as promised.

B. Approximate Transient Analysis

Consider a fixed link and, for $0 \le i \le d$, define $m_t(i)$ to be the probability that a packet i hops from its destination (at the end of the slot) traverses the link during slot t, and set $\boldsymbol{m}_t = (m_t(0), m_t(1), \cdots, m_t(d))$. By assumption, the network is initially empty, so

$$\boldsymbol{m_0} \equiv \mathbf{0}.$$

Given $\boldsymbol{m_t}$ we compute $\boldsymbol{m_{t+1}}$, $t \ge 0$, according to the following *update equations:*

$$m_t = \sum_{i=1}^{d} m_t(i) \tag{3.5}$$

$$\begin{aligned} m_{t+1}(i) = {} & m_t(i-1)p(i-1, m_t, d, v) \\ & + m_t(i+1)(1 - p(i+1, m_t, d, v)) \\ & + \frac{a(m_t, d, v)v}{d} q(i-1)p_o(i-1, m_t, d, v) \\ & + \frac{a(m_t, d, v)v}{d} q(i+1)(1 - p_o(i+1, m_t, d, v)); \\ & 0 \le i \le d \end{aligned} \tag{3.6}$$

where $q(i) = \binom{d}{i}/(2^d - 1)$, and we insist that $p(0, m, d, v) = p_o(0, m, d, v) = 0$ and that

$$q(i) = m_t(i) = p(i, m, d, v) = p_o(i, m, d, v) = 0; \quad i = -1 \quad \text{or} \quad i = d+1.$$

Equation (3.6) is a throughput equation, which follows from the calculations of the previous section. To see this, consider a particular node and note that $dm_{t+1}(i)$ is the expected number of packets that the node transmits during slot $t+1$, such that by the end of slot $t+1$ the packets are at distance i from their destination. Such packets are either

- continuing packets received at the node during slot t with destinations at distance $i-1$ from the node, which are deflected during slot $t+1$,
- continuing packets received at the node during slot t with destinations at distance $i+1$ from the node, which are not deflected during slot $t+1$,
- new packets offered at the node at the beginning of slot $t+1$ with destinations at distance $i-1$ from the node, which are accepted and deflected during slot $t+1$, or

TABLE I

DATA FROM SIMULATIONS AND PREDICTIOJNS FOR A 2^6 NODE HYPERCUBE FOR 25 TIME SLOTS. THE FIRST TWO COLUMNS INDICATE HOW WE VARIED THE OFFERED TRAFFIC RATE v WITH SLOT NUMBER. COLUMNS 3–6 INDICATE, FOR EACH TIME SLOT: THE FRACTION OF THE 384 LINKS USED DURING THE SLOT, THE FRACTION OF OFFERED PACKETS ACCEPTED, THE FRACTION OF PACKETS TRANSITIONS THAT WERE DEFLECTIONS, AND THE MEAN DISTANCE TO DESTINATIONS OF PACKETS CONTINUING AFTER THE END OF THE SLOT. THE PREDICTIONS FOR COLUMNS 3–6 BASED ON THE UPDATE EQUATIONS APPEAR IN COLUMNS 7–10.

		simulation				prediction			
Slot	v	util	accept	deflect	dist	util	accept	deflect	dist
1	6.0	1.0000	1.0000	0.1849	2.4360	1.0000	1.0000	0.1508	2.4874
2	0.0	0.9557	—	0.1826	2.0656	0.9444	—	0.1826	2.1026
3	0.0	0.8333	—	0.1906	1.8228	0.8321	—	0.1838	1.8373
4	0.0	0.6615	—	0.1890	1.6576	0.6659	—	0.1578	1.6305
5	0.0	0.4792	—	0.0652	1.3551	0.4708	—	0.1188	1.4580
6	0.0	0.2786	—	0.1121	1.4091	0.2803	—	0.0766	1.3109
7	0.0	0.1146	—	0.0682	1.1429	0.1307	—	0.0397	1.1907
8	0.0	0.0547	—	0.0000	1.0000	0.0429	—	0.0149	1.0976
9	0.0	0.0078	—	0.0000	1.0000	0.0086	—	0.0032	1.0376
10	0.0	0.0000	—	—	—	0.0009	—	0.0003	1.0097

- new packets offered at the node at the beginning of slot $t+1$ with destinations at distance $i+1$ from the node, which are accepted and not deflected during slot $t+1$.

Now note that the expected number of packets in each of these four groups is d times the corresponding term on the right-hand side of (3.6).

Some statistics that can be approximated using the vector $\boldsymbol{m_t}$, along with their respective approximations, include

- the probability that a link carries a packet during slot t which will continue in the next slot: m_t, defined in (3.5),
- the link utilization at time t : $\sum_{i=0}^{d} m_t(i)$;
- the probabilities of deflection for slot $t+1$: $p(i, m_t, d, v)$ and $p_o(i, m_t, d, v)$ for $0 \le i \le d$;
- the probability of accepting a new packet offered at the beginning of slot $t+1$: $a(m_t, d, v)$;
- the input rate (in packets per node) during slot $t+1$ and the output rate for the previous slot, $va(m_t, d, v)$ and $dm_t(0)$, respectively;
- the distribution of the distance to destination of a typical packet that is deflected in slot $t+1$: the ith term is proportional to

$$m_t(i)p(i, m_t, d, v) + a(m_t, d, v)p_o(i, m_t, d, v)q(i)v/d; \quad 0 \le i \le d.$$

Finally, the rates of convergence of these quantities as $t \to \infty$ are of interest as indications of the rate at which the network approaches equilibrium.

We simulated a hypercube network using one-pass deflection rerouting to compare the transient behavior of the network to predictions derived from the update equations (3.5) and (3.6). Since the update equations were derived using Approximations 3.1 and 3.2 which are not exactly satisfied by the network, the simulations serve to test the accuracy of the approximations. Table I shows the results of one simulation run of a 2^6 node hypercube, together with the corresponding predictions.

The nodes of the network were all filled at the beginning of the first slot, and then new packets were added to the network. The network took ten slots to empty. The fraction of packets deflected was higher in the third and fourth slots than in the first slot, even though there were fewer packets in the network. This can be explained by the fact that after two slots, a typical packet in the network is closer to its destination, and hence has fewer desired outgoing links. The data from simulations is rather close to the predictions, in spite of the fact that we only averaged over the links in the network, an did not average over multiple simulations. The close agreement between simulation and predictions apparent in the data presented in this paper is representative of all the data we have observed.

C. Approximate Steady-State Analysis

We expect that as time t tends to infinity the link utilization vector $\boldsymbol{m_t}$ just defined has a limit $\bar{\boldsymbol{m}} = (\bar{m}(0), \cdots, \bar{m}(d))$. This limit must satisfy the equations obtained by dropping the subscript t from (3.5) and (3.6):

$$\bar{m} = \bar{m}(1) + \ldots + \bar{m}(d) \tag{3.7}$$

$$\begin{aligned}\bar{m}(i) = {} & \bar{m}(i-1)p(i-1, \bar{m}, d, v) \\ & + \bar{m}(i+1)(1 - p(i+1, \bar{m}, d, v)) \\ & + \frac{va(\bar{m}, d, v)}{d}[q(i-1)p_o(i-1, \bar{m}, d, v) \\ & + q(i+1)(1 - p_o(i+1, \bar{m}, d, v))]; \\ & 0 \le i \le d.\end{aligned} \tag{3.8}$$

At this point we are done using Approximations 3.1 and 3.2. Henceforth, we shall assume that $a(m, d, v)$, $p(i, m, d, v)$ and $p_o(i, m, d, v)$ are defined for $0 \le m \le 1$ by (3.1)–(3.4) where U, U_o, V, and V_o have respective binomial distributions $B(d, m)$, $B(d-1, m)$, $B(d, v/d)$, and $B(d-1, v/d)$. The equations (3.7) and (3.8) will be viewed as equations for an unknown vector $\bar{\boldsymbol{m}}$, and the equations will be studied without further reference to the approximations used to derive them. We shall soon show that (3.7) and (3.8) can be reformulated as an equation for the scaler $\bar{m}$.

Summing each side of (3.8) over i with $0 \le i \le d$ yields that

$$a(\bar{m}, d, v)v = \bar{m}(0)d, \tag{3.9}$$

which has the "conservation of mass" interpretation: at equilibrium the rate (in packets per node) at which new packets are

offered and accepted is the same as the rate at which continuing packets are received and absorbed. A rescaled version of (3.8) exposes another useful probabilistic interpretation. If we define $\bar{U} = (\bar{u}(1), \cdots, \bar{u}(d))$ by

$$\bar{u}(i) = \frac{\bar{m}(i)d}{a(\bar{m}, d, v)v} \tag{3.10}$$

then (3.8) becomes

$$\begin{aligned}\bar{u}(i) = \bar{u}(i-1)p(i-1, \bar{m}, d, v) \\ + \bar{u}(i+1)(1 - p(i+1, \bar{m}, d, v)) \\ + q(i-1)p_o(i-1, \bar{m}, d, v) \\ + q(i+1)(1 - p_o(i+1, \bar{m}, d, v)); \qquad 0 \le i \le d\end{aligned} \tag{3.11}$$

We can ascribe a probabilistic interpretation to (3.11) as follows. Let $0 \le m \le 1$ and consider the Markov chain $\boldsymbol{Z}$ with state space $\{0, \cdots, d\}$, initial distribution $(q(0), \cdots, q(d))$, and one-step transition probabilities

$$\begin{aligned}P[Z_1 = i+1 | Z_0 = i] &= p_o(i, m, d, v) \\ P[Z_1 = i-1 | Z_0 = i] &= 1 - p_o(i, m, d, v)\end{aligned} \tag{3.12}$$

and for $t \ge 1$

$$\begin{aligned}P[Z_{t+1} = i+1 | Z_t = i] &= p(i, m, d, v) \\ P[Z_{t+1} = i-1 | Z_t = i] &= 1 - p(i, m, d, v).\end{aligned} \tag{3.13}$$

Since $p(0, \cdot, \cdot, \cdot) = p_o(0, \cdot, \cdot, \cdot) = 0$, state 0 is absorbing. We can think of Z_t as a description of the time evolution of the distance of a typical packet from its destination, when the mean number of continuing packets at each node in each slot is m. Define $T(m, d, v)$ by

$$T(m, d, v) = E[\min\{t : Z_t = 0\}]. \tag{3.14}$$

Equation (3.11) implies that $\bar{u}(i)$ is the expected number of visits of Z to state i when the parameter m for Z is set equal to $\bar{m}$:

$$\bar{u}(i) = E[\text{number of } t \ge 1 \text{ such that } Z_t = i] \quad \text{for } m = \bar{m}. \tag{3.15}$$

Using (3.15) and (3.10) we see that

$$T(\bar{m}, d, v) - 1 = \sum_{i=1}^{d} \bar{u}(i) \frac{\bar{m}d}{a(\bar{m}, d, v)v} \tag{3.16}$$

or

$$\bar{m} = \frac{(T(\bar{m}, d, v) - 1)a(\bar{m}, d, v)v}{d}. \tag{3.17}$$

We refer to equation (3.17) as the *fixed point* equation for $\bar{m}$. This equation is a version of Little's law (though applied to our approximations): packets are accepted at rate $a(\bar{m}, d, v)v/d$ packets per link per slot, and each packet spends, on average, $T(\bar{m}, d, v) - 1$ slots in the network, not counting the last slot.

We've shown that if $\bar{\boldsymbol{m}}$ satisfies (3.8), then $\bar{m}$ satisfies (3.17). Conversely, suppose $\bar{m}$ is a solution to (3.17). Then $\bar{m}$ determines the Markov chain $\boldsymbol{Z}$ with $m = \bar{m}$ via (3.12)

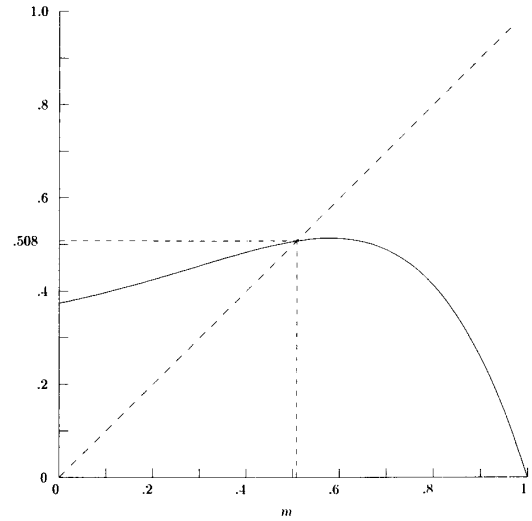


Fig. 1. $(T(m, v, d) - 1)\, a(m, v, d)v/d$ versus m for $d = 6$, $v = 1$. Here $\bar{m} = 0.5080596$..

and (3.13), which determines $\bar{u}(i)$ via (3.15), which in turn determines $\bar{m}(i)$ via (3.10), and so determines $\bar{\boldsymbol{m}}$. Equivalently, $\bar{m}$ determines $\bar{u}(i)$ for $1 \le i \le d$ via (3.11)–(3.13) with $m = \bar{m}$. Hence, there is a one-to-one correspondence between solutions $\bar{\boldsymbol{m}}$ to (3.8) and solutions $\bar{m}$ to (3.17).

The right-hand side of (3.17) is continuous in $\bar{m}$, positive at $\bar{m} = 0$, and zero at $\bar{m} = 1$ (Lemma 3.2 below shows that $T(m, d, v)$ is bounded by a function of d alone), so a solution $\bar{m}$ exists, and is easy to find. Numerical experiments have led us to believe that (3.17) has a unique solution $\bar{m}$, though we have not proved this. See Fig. 1. Given a solution $\bar{m}$ we can calculate estimates of the steady-state counterparts for the statistics mentioned in the previous section. In addition we can use the Markov chain interpretation $\boldsymbol{Z}$ to calculate statistics of packet delay. The quantity $T(\bar{m}, d, v)$ represents the average packet delay. Formulae similar to (3.11) and (3.16) can be used to obtain higher moments. Standard absorbing Markov chain methods can be used to calculate, for example, the probability that the delay exceeds a given duration.

Simulation results and predictions based on the fixed point equations are given in Tables II and III, respectively. Each line in Table II corresponds to a network simulated for 1100 time slots, with the reported averages coming from the final 1000 time slots. Table IV indicates the distribution of how far a typical deflection is from the destination of the deflected packet. The entries in the "predicted" columns are obtained by normalizing $\bar{m}(i)p(i, \bar{m}, d, v) + a(\bar{m}, d, v)vp_o(i, \bar{m}, d, v)q(i)/d$. Thus, for example, the entry 0.5413 suggests that about 54% of the time there is a deflection, the deflected packet is one hop from its destination.

D. Asymptotics of Steady State Approximations

In this section we give the asymptotics of the key performance statistics, $\bar{m}$, $a(\bar{m}, d, v)$, and $T(\bar{m}, d, v)$, as the hypercube dimension $d \to \infty$. Proofs may be found in the appendix. Here $\bar{m}$ is defined to be a solution to the fixed point

TABLE II
AVERAGE LINK UTILIZATION, FRACTION OF PACKETS ACCEPTED, DELAY OF ACCEPTED PACKETS, AND FRACTION OF TRANSITIONS THAT WERE DEFLECTIONS, FOR A 1000 TIME SLOT SIMULATION OF A 2^6 NODE HYPERCUBE NETWORK

v	link util.	accept frac.	delay	def. frac.
0.2	0.1048	1.0000	3.1633	0.0166
0.4	0.2218	1.0000	3.2959	0.0379
0.6	0.3504	0.9975	3.5017	0.0647
0.8	0.5029	0.9822	3.8297	0.1014
1.0	0.6512	0.9338	4.2092	0.1383
1.2	0.7766	0.8332	4.6794	0.1740
1.4	0.8516	0.7303	4.9804	0.1945
1.6	0.8942	0.6463	5.1923	0.2059
1.8	0.9215	0.5819	5.3018	0.2129
2.0	0.9427	0.5205	5.4421	0.2203
2.2	0.9575	0.4739	5.5056	0.2234
2.4	0.9672	0.4351	5.5689	0.2265
2.6	0.9754	0.4011	5.6127	0.2287
2.8	0.9809	0.3717	5.6632	0.2310
3.0	0.9856	0.3453	5.6956	0.2323

TABLE III
AVERAGE LINK UTILIZATION, FRACTION OF PACKETS ACCEPTED, AVERAGE DELAY OF ACCEPTED PACKETS, AND FRACTION OF TRANSITIONS THAT ARE DEFLECTIONS, *PREDICTED* FOR A 2^6 NODE HYPERCUBE NETWORK

v	link util.	accept frac.	delay	def. frac.
0.2	0.1051	1.0000	3.1525	0.0166
0.4	0.2197	0.9998	3.2967	0.0378
0.6	0.3499	0.9975	3.5074	0.0655
0.8	0.5022	0.9832	3.8310	0.1022
1.0	0.6629	0.9289	4.2816	0.1441
1.2	0.7827	0.8307	4.7112	0.1766
1.4	0.8538	0.7295	5.0159	0.1962
1.6	0.8968	0.6440	5.2226	0.2082
1.8	0.9248	0.5743	5.3675	0.2161
2.0	0.9441	0.5175	5.4726	0.2216
2.2	0.9579	0.4706	5.5508	0.2255
2.4	0.9681	0.4314	5.6102	0.2284
2.6	0.9758	0.3981	5.6558	0.2306
2.8	0.9816	0.3696	5.6912	0.2323
3.0	0.9861	0.3449	5.7188	0.2335

equation (3.17), or equivalently $\bar{\boldsymbol{m}}$ is defined to be a solution to (3.7) and (3.8). We also report some numerical comparisons between these quantities for finite (small) d and their limits.

The following lemma bounds the deflection probabilities $p(i,m,d,v)$ and $p_o(i,m,d,v)$ and the average delay $T(m,d,v)$, uniformly in the free parameter m.

Lemma 3.1: For all $1 \le i \le d, 0 \le m \le 1, v \ge 0$, and $d \ge 1$,

$$p(i,m,d,v) \le \frac{1}{i+1}, \tag{3.18}$$

$$p_o(i,m,d,v) \le \frac{1}{i+1} \tag{3.19}$$

and

$$\frac{d}{2} \le T(m,d,v) \le \frac{d}{2} + 6 + 2\log d. \tag{3.20}$$

These bounds help to explain why the routing works well, not only at steady state, but also at all times $t \ge 0$. The first two bounds state that the probability that a packet is deflected falls off inverse linearly with its distance from its destination, independently of all other parameters of the model. Thus, a packet in transit in the network is strongly attracted to its destination. The third bound states that the packet's average delay differs from the minimal value $\sim d/2$ by at most $O(\log d)$, independently of all other model parameters. (In practice d is small; d is the log to base two of the number of nodes in the network.) This means that on average a packet suffers at most $O(\log d)$ deflections. Finally, the constant 6 in (3.20) can be improved to 4, with more work.

TABLE IV
THE DISTRIBUTION OF HOW FAR A TYPICAL DEFLECTION OCCURS FROM THE DESTINATION OF THE DEFLECTED PACKET. THE PREDICTIONS ARE BASED ON THE SOLUTION OF THE FIXED POINT EQUATION (3.17) AND THE SIMULATION DATA SHOW THE EMPIRICAL DISTRIBUTION FOR A 1000 TIME SLOT RUN FOR $d = 6$ AND A 200 TIME SLOT RUN FOR $d = 8$

	$d = 6, v = 2$		$d = 8, v = 2$	
distance	prediction	simulation	prediction	simulation
1	0.5413	0.5332	0.4654	0.4648
2	0.3259	0.3305	0.3240	0.3195
3	0.1070	0.1087	0.1403	0.1424
4	0.0233	0.0245	0.0514	0.0531
5	0.0026	0.0031	0.0155	0.0166
6	0.0000	0.0000	0.0032	0.0033
7	—	—	0.0003	0.0003
8	—	—	0.0000	0.0000

In our model a packet's destination is on average slightly more than $d/2$ hops from its source. A node can transmit at most d packets per unit time. Therefore, the throughput in packets per node cannot exceed $d/(d/2) = 2$. Our next result implies that the throughput $a(\bar{m},d,v)v \to 2 \wedge v$ as $d \to \infty$ for v fixed. This is optimal in the sense that (asymptotically) no packets are blocked if $v < 2$ and *well behaved at saturation* in the sense that packets flow into the network at the maximal rate if the offered load $v \ge 2$.

Theorem 3.1: For all $v \ge 0$ and $d \ge 1$,

a) there exists a solution $\bar{\boldsymbol{m}}$ to (3.7) and (3.8); equivalently there exists a solution $\bar{m}$ to (3.17), and
b) for v fixed as $d \to \infty$,

$$\bar{m} \to \frac{v}{2} \wedge 1 \tag{3.21}$$

$$a(\bar{m},d,v) \to \frac{v}{2} \wedge 1 \tag{3.22}$$

$$p(i,\bar{m},d,v) \to p^{\infty}(i,v); \qquad i \ge 0 \tag{3.23}$$

where

$$p^{\infty}(i,v) = \frac{\left(\frac{v}{2} \wedge 1\right)^i}{i+1}. \tag{3.24}$$

Fig. 2 depicts $a(\bar{m},d,v)$ plotted against d for $v = 1.5$. Unfortunately, if v is near 2 then for practical values of d (say $d \le 20$), the limiting value 1.0 is only a rough approximation to the true one.

The next theorem gives more precise results on how full the network is when $v \ge 2$ and the network is moderately large.

Theorem 3.2: If $v \ge 2$ then

$$\bar{m} = 1 - \frac{G_v^{-1}(2)}{d} + o(1/d); \qquad \text{as } d \to \infty \tag{3.25}$$

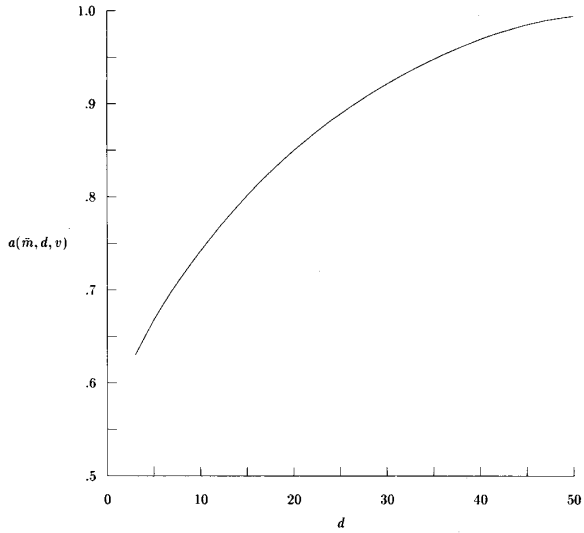


Fig. 2. $a(\bar{m}, d, v)$ versus d for $v = 1.5$.

where $G_x^{-1}(y)$ is the inverse of the function $G_x(y) = E[X \wedge Y]$ where X and Y represent independent, Poisson distributed random variables with respective means x and y.

Finally, we can characterize the asymptotics of the average packet delay $T(\bar{m}, d, v)$ for $v < 2$.

Theorem 3.3: For $0 \leq v \leq 2$ fixed,

$$T(\bar{m}, d, v) - T^{\infty}(d, v) \to 0, \qquad \text{as } d \to \infty \tag{3.26}$$

where

$$T^{\infty}(d, v) = \frac{d}{2(1 - 2^{-d})} + 2C\left(\frac{v}{2}\right) \tag{3.27}$$

$$C(\mu) = \sum_{1 \leq k \leq j \leq \infty} \prod_{i=k}^{j} \frac{\mu^i}{1 + i - \mu^i}. \tag{3.28}$$

For $0 \leq \mu < 1$,

$$-\frac{\log(1-\mu)}{\mu} - 1 \leq C(\mu) \leq \frac{\log(1-\mu)}{\mu} + 1.7. \tag{3.29}$$

The quantity $C(v/2)$ represents the expected number of deflections a typical packet suffers at equilibrium, in the limit as $d \to \infty$. This quantity diverges for $v \geq 2$. Remarkably, it remains infinite for $v < 2$. Moreover, the bounds (3.29) show that $C(v/2)$ grows very slowly as $v \to 2$ from below. This further illustrated in Table V. The bound for $p(i, \bar{m}, d, v)$ at the end of the proof of Lemma 5.5 implies that if $v < 2$ then every moment of the number of deflections remains bounded as $d \to \infty$. If $v \geq 2$ then the average number of deflections tends to ∞ as $d \to \infty$.

Table VI compares the average delay observed in simulations to the predicted value $T(\bar{m}, d, v)$, and to the asymptotic approximation $T^{\infty}(d, v)$, for the different size networks. Each simulation result in this table was collected by averaging over the last 10^4 steps of a run of $2 \cdot 10^4$ steps. The simulation for $d = 13$ entailed simulating 106,496 nodes for 20 000 time slots, and took over 20 h on a MIPS RS2000 high-performance workstation, whereas the apparently accurate approximation $T(\bar{m}, d, v)$ took less than ten seconds to compute.

TABLE V
SOME VALUES OF $C(\mu)$

μ	$C(\mu)$
0.2	0.117
0.4	0.288
0.6	0.573
0.8	1.15
0.9	1.82
0.99	4.18
0.999	6.50
0.9999	8.81
0.99999	11.11

TABLE VI
$T_{\text{sim}}(d, v)$ [AVERAGE DELAY MEASURED IN SIMULATION] $T(\bar{m}, d, v)$ AND $T^{\infty}(d, v)$ VERSUS d, FOR $v = 1$

d	$T_{\text{sim}}(d, 1)$	$T(\bar{m}, d, 1)$	$T^{\infty}(d, 1)$
2	1.74	1.805	2.155
3	2.46	2.491	2.536
4	3.09	3.119	2.955
5	3.70	3.713	3.402
6	4.30	4.282	3.869
7	4.84	4.826	4.349
8	5.37	5.349	4.837
9	5.87	5.853	5.331
10	6.36	6.343	5.827
11	6.84	5.826	6.324
12	7.32	7.304	6.823
13	7.79	7.782	7.323
14	—	8.261	7.822
15	—	8.741	8.322
16	—	9.224	8.822
17	—	9.709	9.322
18	—	10.195	9.822
19	—	10.683	10.322
20	—	11.172	10.822

IV. INPUT QUEUEING AND FINAL REMARKS

We have presented a model and a corresponding approximate analysis of deflection routing in hypercube networks. Monte–Carlo simulations showed that the analytic approximations are quite accurate. In the model it is assumed that the traffic is statistically uniform, in the sense that the nodes are statistically indentical in packet generation and the destination of each packet is equally likely to be any node excluding the packet's source.

Using the results of our analysis, the delay and throughput of the network can be computed easily, for finite time intervals and at equilibrium. We found that performance is quite good. An asymptotic analysis of our performance predictions showed that as the network size increases performance becomes optimal.

- The component of delay due to deflection is minimal.
- The achieved throughput equals the minimum of the rate at which packets are offered and the raw network capacity.

The conclusions reached by Syzmanski [13] are consistent with ours. He found that deflection routing works well in hypercube networks.

In the model we described in Section II-A and subsequently analyzed, if the number of new arrivals at a node plus the number of continuing packets at the node during a time slot is greater than d, then some of the arrivals are simply dropped. An alternative procedure would be to queue the excess arrivals until they can be admitted into the network. Such input queueing results in a possible queueing delay for packets as they enter the network. We shall briefly explain how our approximate analysis extends to analyze the network with input queues.

If at a node at the beginning of a slot there are U continuing packets, Q packets in the node's input queue and A new arrivals, then $\min(Q+A, d-U)$ new packets enter the network. Assume that the number of new packets that arrive at a node during a time slot has a Poisson distribution with parameter λ where $\lambda > 0$, and that the numbers of arrivals from node to node and slot to slot are independent. If the system is stable then the steady-state throughput will also be λ. We find a value[2] of offered traffic v so that the throughput $va(\bar{m}, d, v)$ predicted by the model of Sec tion III-C is equal to λ. By that model, the number of new packets that can be potentially accepted at a node in a time slot has the binomial distribution $B(d, 1-\bar{m})$. By invoking the additional approximation that the numbers of packets that can be potentially accepted from slot to slot are independent, we model the queue of packets waiting at a node as a discrete-time queue where the number of arrivals in each slot is Poisson with mean λ and the number of potential services in each slot has the binomial distribution $B(d, 1-\bar{m})$.

This batch queueing model can be approximated by, and the mean waiting time in queue bounded by, that of a second, simpler queueing model. The second queueing model operates using d mini-slots per slot where the number of arrivals in any mini-slot has the Poisson distribution with mean λ/d and the number of potential departures during any mini-slot is one with probability $1-\bar{m}$, and zero otherwise. The second system is similar to the first, except that potential services are offered at times spread through a slot rather than just at the end of the slot. As a result the mean number of customers held over after a typical slot in the first system is less than or equal to the mean number held over after a minislot in the second system, which is given by

$$\bar{X} = \frac{(\lambda/d)^2 + 2\bar{m}\lambda/d}{2(1-\bar{m}-\lambda/d)}. \tag{4.1}$$

Thus the mean waiting time of customers in the first queueing model is less than or equal to $\bar{X}/\lambda$, and, further thought shows, greater than or equal to $\bar{X}/\lambda - 1$. Thus, we expect $\bar{X}/\lambda$ to serve as a good approximation to the waiting time in a hypercube network with Poisson arrivals and input queueing.

Note that input queueing may affect the validity of Approximations 3.1 and 3.2. A comparison with simulations is indicated in Fig. 3 for a 64 node hypercube network. As in other simulations reported in this paper, packets are generated at random and traced through the network. The good agreement indicated in Fig. 3 gives further evidence that Approximations 3.1 and 3.2 are valid. In addition, even for arrival rates λ up to 95% of the maximum that can be tolerated, the mean input queueing delay (lower curve) is small compared to the mean transit delay (upper curve) within the network. This can probably be attributed to the fact that the multiple outgoing links at a node make it resemble a multiserver queue.

[2]Numerically we observe that the throughput is monotone increasing in v, except for a slight (0.2%) decrease beyond a certain value of v. Thus, λ uniquely determines v unless λ is very near the maximum possible throughput.

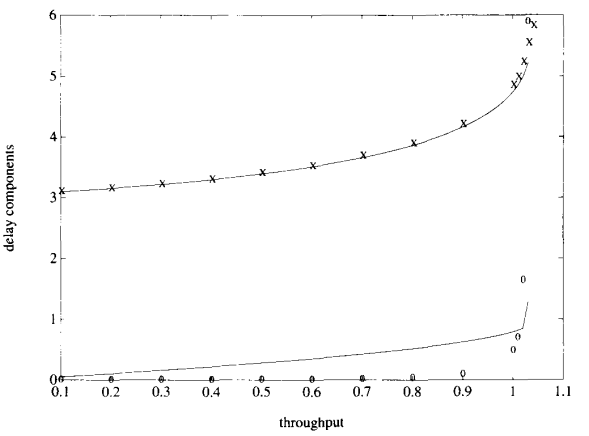


Fig. 3. A packet's delay is the sum of its transit delay (in the network), and its queueing delay (in the input queue at the node where the packet originates). The x's mark the average transit delay and the o's the average queueing delay, resp., measured in simulation, for a 64 node hypercube network and a range of arrival rates λ. The upper and lower curves show the corresponding analytic predictions.

The most intriguing open question about deflection routing is perhaps how does the algorithm perform if the traffic pattern is not uniform. It is important to study the behavior of the network under unbalanced loads. See [4] for some bounds for the worst case. An analysis similar to Valiant's analysis of randomized routing algorithms [15], [16] would be very valuable. Intuitively, deflection routing should overcome the hot spots that might arise in the network in the important case where the traffic is not uniform, but is balanced in the sense that the total load directed out of and into each node is the same. In this vein, we note that the deflection routing algorithm can easily be extended to use buffers internal to the nodes (c.f. [6]), and doing so may increase throughput under uniform traffic. However, additional buffering lessens the algorithm's adaptivity to congestion, and may be detrimental to performance if the traffic is not uniform.

Appendix

The purpose of this appendix is to prove the results stated in Section III-D. Let Θ denote a discrete time Markov chain with transition probabilities $\pi = (\pi_0, \pi_1, \pi_2 \cdots)$ where

$$\begin{aligned}\pi_i &= \Pr(\Theta_{t+1} = i+1 | \Theta_t = i) \\ &= 1 - \Pr(\Theta_{t+1} = i-1 | \Theta_t = i); \qquad i \geq 0.\end{aligned}$$

Define

$$D(\pi) = \min\{t \geq 0 : \Theta_t = 0\}$$

and, for any integer $n \geq 0$, let $E_n D(\pi)$ denote the expected value of $D(\pi)$ if $\Theta_0 = n$. Moreover, for any probability distribution $\boldsymbol{q} = (q_0, q_1, q_2, \cdots)$ on the nonnegative integers, let $E_{\boldsymbol{q}} D(\pi)$ denote the expected value of $D(\pi)$ if $\Theta_o = i$ with probability q_i.

Lemma 5.1: For all $n \geq 0$, $E_n D(\pi) = n + \sum_{k=1}^{n} U_k$ where

$$U_k = 2 \sum_{j=k}^{\infty} \prod_{i=k}^{j} \frac{\pi_i}{1 - \pi_i}. \tag{5.1}$$

Proof: Simplifying notation, let $\tau(n) = E_n D(\pi)$. First, choose $N > 1$ and suppose that $\pi_i = 0$ for $i \geq N$. Then $(\tau(n) : n \geq 1)$ is the unique solution to

$$\tau(n) = 1 + \pi_n \tau(n+1) + (1 - \pi_n)\tau(n-1); \qquad n \geq 1$$

where $\tau(0) = 0$. It is easy to check that the formula for $\tau(n) = E_n D(\pi)$ given in the Lemma satisfies this equation. This proves the Lemma if $\pi_i = 0$ for $i \geq N$. (Another, more revealing proof, is obtained by noting that $\prod_{i=k}^{j} \pi_i/(1 - \pi_i)$ is the mean number of transitions from j to $j+1$ that Θ makes before $k - 1$ is reached, starting from k, and $U_k + 1$ is the mean time to hit $k - 1$, starting from k).

Turn, now, to the general case of the Lemma. Fix $n \geq 1$ and for $N > n$ let $\pi_i^N = \pi_i$ for $i < N$ and $\pi_i^N = 0$ for $i \geq N$. Write $\tau^N(n) = E_n D(\pi^N)$. Then by the monotone convergence theorem,

$$\begin{aligned} \tau(n) &= \lim_{N \to \infty} E_n \min\{t \geq 0 : \Theta_t = 0 \text{ or } \Theta_t = N\} \\ &\leq \lim_{N \to \infty} \tau^N(n) \leq \tau(n). \end{aligned}$$

Hence, $\lim_{N \to \infty} \tau^N(n) = \tau(n)$. Since $\pi_i^N = 0$ for $i \geq N$, the first part of this proof implies that

$$\tau^N(n) - n = 2 \sum_{k=1}^{n} \sum_{j=k}^{N-1} \prod_{i=k}^{j} \frac{\pi_i}{1 - \pi_i}. \tag{5.2}$$

As N tends to infinity, the right-hand side of (5.2) converges to U_k by the monotone convergence theorem, and the Lemma is proved. □

Proof of Lemma 3.1: We will use (3.2)–(3.4). For $0 \leq k \leq d - 1$ and $1 \leq i \leq d$,

$$\begin{aligned} H(k, d, i) &\leq \frac{1}{k+1} \sum_{j=i}^{k} \left(\frac{j}{d}\right)^i \leq \frac{1}{k+1} \int_i^{k+1} \left(\frac{x}{d}\right)^i dx \\ &\leq \frac{1}{i+1} \left(\frac{k+1}{d}\right)^i \leq \frac{1}{i+1}, \end{aligned} \tag{5.3}$$

which establishes (3.18). Using (3.4), (5.3), and the fact that $a(m, d, v) = E\left[\frac{(1+V_o)\wedge(d-U)}{1+V_o}\right]$ establishes (3.19).

Since $Z_t \geq Z_0 - t$ for all t, we see that $T(m, d, v) \geq E[Z_0] = d/(2(1 - 2^{-d})) \geq d/2$, which proves the first inequality of (3.20). If we choose $\pi_i = 1/(i+1)$ and let the Markov chain Θ of Lemma 5.1 have the initial distribution $\boldsymbol{q}$ given by (2.1), then (3.18) implies that $\boldsymbol{Z}$ is stochastically dominated by Θ. Thus $T(m, d, v) \leq E_{\boldsymbol{q}} D(\pi)$. Now

$$\begin{aligned} U_k &= 2\left(\frac{1}{k} + \frac{1}{k}\frac{1}{k+1} + \frac{1}{k}\frac{1}{k+1}\frac{1}{k+2} + \cdots\right) \\ &\leq \frac{2}{k}\left(1 + \frac{1}{k+1} + \left(\frac{1}{k+1}\right)^2 + \cdots\right) \\ &\leq \frac{2k+2}{k^2}. \end{aligned}$$

Thus, by Lemma 5.1,

$$\begin{aligned} E_{\boldsymbol{q}} D(\pi) &= \sum_{n=1}^{d} q_n \left\{ n + \sum_{k=1}^{n} U_k \right\} \leq \frac{d}{2(1 - 2^{-d})} + \sum_{k=1}^{d} U_k \\ &\leq \frac{d}{2(1 - 2^{-d})} + 4 + \int_1^d \frac{2x+2}{x^2}\, dx \\ &\leq \frac{d}{2} + 6 + 2\log d \end{aligned}$$

as was to be shown. □

It turns out that the deflection and acceptance probabilities tend to simple limits as $d \to \infty$. Let

$$G_x(y) = E[X \wedge Y] \tag{5.4}$$

where X and Y are independent, Poisson distributed random variables with respective means x and y.

Lemma 5.2: For all $i \geq 0$ and $v \geq 0$, as $d \to \infty$,

$$p(i, m, d, v) \to \frac{m^i}{i+1}, \quad \text{uniformly for } 0 \leq m \leq 1; \tag{5.5}$$

$$a(m, d, v)v - G_v(d(1 - m)) \to 0, \quad \text{uniformly for } 0 \leq m \leq 1. \tag{5.6}$$

Moreover, for any $0 \leq m_0 < 1$,

$$a(m, d, v) \to 1, \qquad \text{uniformly for } 0 \leq m \leq m_0. \tag{5.7}$$

Proof: We again concentrate on (3.2). By (5.3) we see that for $0 \leq k \leq d - 1$ and $1 \leq i \leq d$,

$$H(k, d, i) \leq \frac{1}{i+1}\left(\frac{k+1}{d}\right)^i \tag{5.8}$$

We also have

$$\begin{aligned} H(k, d, i) &\geq \frac{1}{k+1} \sum_{j=i}^{k} \left(\frac{j-i}{d}\right)^i \geq \frac{1}{k+1} \int_i^k \left(\frac{x-i}{d}\right)^i dx \\ &= \frac{(k-i)^{i+1}}{(i+1)d^i(k+1)}. \end{aligned} \tag{5.9}$$

Bounds (5.8) and (5.9) imply that for i fixed,

$$\left| H(k, d, i) - \frac{1}{i+1}\left(\frac{k}{d}\right)^i \right| \to 0 \tag{5.10}$$

as $d \to \infty$, uniformly as k varies with $0 \le k \le d-1$. As $d \to \infty$,

$$\frac{E[(U_o + V + 1) \wedge d]}{d} - m \to 0 \quad \text{and}$$
$$\frac{\mathrm{Var}((U_o + V + 1) \wedge d)^{1/2}}{d} \to 0.$$

Hence, the uniform convergence in (5.10) and the fact that $m^i/(i+1)$ is uniformly continuous in m imply (5.5) of the Lemma.

Next, recall the definition (3.1) of $a(m,d,v)$ and let $\tilde{U} = d - U$. Then $\tilde{U}$ has the binomial distribution $B(d, 1-m)$ and

$$a(m,d,v)v = E\left[\tilde{U} \wedge V\right] = \sum_{i=0}^{d}\sum_{j=0}^{i-1} jP\left[\tilde{U} = j\right]P[V = i] + \sum_{i=0}^{d} iP\left[\tilde{U} \ge i\right]P[V = i]. \tag{5.11}$$

Let $Poi(x)$ denote a Poisson distribution random variable with mean x. By the well known approximation of the binomial distribution by the Poisson distribution we find that for each i and j fixed, uniformly for m, $P[V = i] \to P[Poi(v) = i]$, $P[\tilde{U} = j] \to P[Poi(d(1-m)) = j]$ and $P[\tilde{U} \ge i] \to P[Poi(d(1-m)) \ge i]$ as $d \to \infty$. Also,

$$\begin{aligned} P[V = i]\frac{i!}{v^i} &= \left(1 - \frac{v}{d}\right)^d \frac{d_i}{(d-v)^i} \\ &\le \left(1 - \frac{v}{d}\right)^d \prod_{1 \le j \le u \wedge \lfloor v \rfloor} \left(\frac{d-j}{d-v}\right) \\ &\le \left(1 - \frac{v}{d}\right)^{d-v} \le \exp\left(-v + \frac{v^2}{d}\right) \end{aligned}$$

so that $P[V = i] \le P[Poi(v) = i]\exp(v^2/d)$. Thus, for all d with $d \ge v^2$, the terms on the right-hand side of (5.11) a term-by-term less than the terms in the following convergent series:

$$\sum_{i=0}^{\infty}\sum_{j=0}^{i-1} jP[Poi(v) = i]e + \sum_{i=0}^{\infty} iP[Poi(v) = i]e.$$

Thus, the dominated convergence theorem implies that the limit can be taken in (5.11) term-by-term, so as $d \to \infty$, $a(m,d,v)v$ converges uniformly for $0 \le m \le 1$ to

$$\sum_{i=0}^{\infty}\sum_{j=0}^{i-1} jP[Poi(d(1-m)) = j]P[Poi(v) = i] + \sum_{i=0}^{\infty} iP[Poi(d(1-m)) \ge i]P[Poi(v) = i]$$

which is equal to $G_v(d(1-m))$. This proves part (5.6). Equation (5.7) is a consequence. □

Proof of Theorem 3.1: As mentioned in Section III-C, part (a) follows from the fact that the right-hand side of (3.17) is continuous in $\overline{m}$, positive at $\overline{m} = 0$, and zero at $\overline{m} = 1$. The equivalence referred to in part (a) was also explained in Section III-C. For part (b), suppose v is fixed and d allowed to increase. Equation (3.21) follows from (3.17) and the uniform bounds on $a(m,d,v)$ and $T(m,d,v)$, provided in (5.7) and (3.20), respectively. The fixed point equation (3.17), the equation (3.21) for $\overline{m}$ just established, and the bound (3.20) for $T(m,d,v)$ imply (3.22). Equation (3.23) for the deflection probabilities follows from the equation (3.21) for $\overline{m}$ and (5.5) on the asymptotics of the deflection probabilities, for any m. □

Proof of Theorem 3.2: If $v \ge 2$ then $a(\overline{m},d,v)v \to 2$ as $d \to \infty$ by Theorem 3.1, (3.22) on $a(\overline{m},d,v)$. By (5.6) of Lemma 5.2 it follows that $G_v(d(1-\overline{m})) \to 2$, which, since G_v is a continuous increasing function, implies that $d(1-\overline{m}) \to G_v^{-1}(2)$. □

To prove Theorem 3.3 we shall compare the underlying Markov chain $\boldsymbol{Z} = \boldsymbol{\Theta}(\boldsymbol{p}(\overline{m},d,v))$ to the Markov chain $\boldsymbol{\Theta}(\boldsymbol{p}^\infty(v))$ where $\boldsymbol{\Theta}(\pi)$ denotes the Markov chain discussed at the beginning of this section, with probabilities of steps to the right given by the vector π, and $\boldsymbol{p}(\overline{m},d,v) = (p(i,\overline{m},d,v) : i \ge 0)$ $\boldsymbol{p}^\infty(v) = (p^\infty(i,v) : i \ge 0)$. Recall the notation

$$D(\pi) = \min\{t \ge 0 : \Theta_t = 0\}$$

and recall that, for any integer $n \ge 0$, $E_n D(\pi)$ denotes the expected value of $D(\pi)$ if $\Theta_0 = n$. For any probability distribution $\boldsymbol{r} = (r_0, r_1, r_2, \cdots)$ on the nonnegative integers, let $E_{\boldsymbol{r}} D(\pi)$ denote the expected value of $D(\pi)$ if $\Theta_0 = i$ with probability r_i. Thus, $T(\overline{m},d,v) = E_{\boldsymbol{q}}\boldsymbol{\Theta}(\boldsymbol{p}(\overline{m},d,v))$ where $\boldsymbol{q} = (q_i : i \ge 0)$ is as defined in (2.1).

Lemma 5.3: As $d \to \infty$, $T(\overline{m},d,v) - E_{\boldsymbol{q}}D(\boldsymbol{p}(\overline{m},d,v)) \to 0$.

Proof: Both $T(\overline{m},d,v)$ and $E_{\boldsymbol{q}}D(\boldsymbol{p}(\overline{m},d,v))$ represent the mean time to hit zero for a Markov random walk. The only difference is in the deflection probability for the first step of the walk, which in both cases tends to zero as $d \to \infty$. Details are left to the reader. □

Define the class

$$\boldsymbol{\Pi} = \{\boldsymbol{p} = (p(0), p(1), p(2), \cdots) : p(i) \le 1/(i+1), i \ge 0\}$$

and the metric

$$\|\boldsymbol{p} - \boldsymbol{p}'\| = \sum_{i=0}^{\infty} |p(i) - p'(i)|; \qquad \boldsymbol{p}, \boldsymbol{p}' \in \boldsymbol{\Pi}.$$

Lemma 5.4: For any $\boldsymbol{p}$ and $\boldsymbol{p}'$ in $\boldsymbol{\Pi}$ and any probability distribution $\boldsymbol{q}$ on the nonnegative integers,

$$\left|E_{\boldsymbol{q}}D(\boldsymbol{p}) - E_{\boldsymbol{q}}D(\boldsymbol{p}')\right| \le 16\|\boldsymbol{p} - \boldsymbol{p}'\|.$$

Proof: It suffices to prove that for p in $\boldsymbol{\Pi}$ and $n, n' \ge 1$ that $0 \le (\partial E_{n'} D(\boldsymbol{p})/\partial[p_n]) \le 16$. By Lemma 5.1,

$$\begin{aligned} 0 \le \frac{\partial E_{n'} D(\boldsymbol{p})}{\partial[p_n]} &\le \sum_{k=1}^{\infty} \frac{\partial U_k}{\partial[p_n]} \\ &= 2\sum_{k=1}^{n}\sum_{j=n}^{\infty} \frac{\partial}{\partial[p_n]} \prod_{i=k}^{j} \frac{p_i}{1-p_i} = \frac{2A_nB_n}{(1-p_n)^2} \end{aligned}$$

where $A_n = 1 + \sum_{k=1}^{n-1} \prod_{i=k}^{n-1} p_i/(1-p_i)$ and $B_n = 1 + \sum_{j=n+1}^{\infty} \prod_{i=n+1}^{j} p_i/(1-p_i)$. Now $1/(1-p_n)^2 \leq (1+1/n)^2$. Note that $A_1 = 1$ and $A_{n+1} = 1 + (p_n A_n)/(1-p_n) \leq 1 + A_n/n$ for all n so that $A_n \leq 2$ for all $n \geq 2$. Finally,

$$B_n \leq 1 + \frac{1}{n+1} + \frac{1}{(n+1)^2} + \cdots = \frac{n+1}{n}.$$

Thus, $(2A_1B_1)/(1-p_1)^2 \leq (2)(1)(2)/(0.5)^2 = 16$ and for $n \geq 2$, $(2A_nB_n)/(1-p_n)^2 \leq 2(1+1/n)^3 \leq 16$. □

Lemma 5.5: For $0 \leq v < 2$ fixed,

$$\|\boldsymbol{p}(\bar{m}, d, v) - \boldsymbol{p}^{\infty}(v)\| \to 0 \quad \text{as} \quad d \to \infty.$$

Proof: By (5.5) of Lemma 5.2 and (3.21) of Theorem 3.1, $p(i, \bar{m}, d, v) \to p^{\infty}(i, v)$ for fixed i. By the dominated convergence theorem, it is thus sufficient to find a summable (over all i) upper bound on $p(i, \bar{m}, d, v)$ that holds for all sufficiently large d. By the definition of $p(i, m, d, v)$ [(3.2)] and the bound (5.8), we see that $p(i, \bar{m}, d, v) \leq E[W^i]$ for $1 \leq i \leq d$ where $W = \frac{(U_o+V+1)\wedge d}{d}$. By Cauchy's inequality and the convention that $p(i, \bar{m}, d, v) = 0$ for $i > d$, we have $p(i, \bar{m}, d, v) \leq \beta(d)^i$ for all $i \geq 1$ where $\beta(d) = E[W^d]^{1/d}$. By applying the Chernoff bound we can show that there is a constant d_o and a constant $\alpha_o(v) < 1$ so that $\beta(d) \leq \alpha_o(v)$ for all d with $d \geq d_o$. Details are omitted. Hence, $p(i, \bar{m}, d, v) \leq \alpha(v)^i$ for all i and all d with $d \geq d_o$. □

Proof of Theorem 3.3: Lemmas 5.4 and 5.5 together imply that

$$\left|E_{\boldsymbol{q}} D(\boldsymbol{p}(\bar{m}, d, v)) - E_{\boldsymbol{q}} D(\boldsymbol{p}^{\infty}(v))\right| \to 0.$$

By Lemma 5.3, it suffices to show that as $d \to \infty$,

$$\left|E_{\boldsymbol{q}} D(\boldsymbol{p}^{\infty}(v)) - \frac{d}{2(1-2^{-d})} - 2C(v/2)\right| \to 0.$$

By definition,

$$E_{\boldsymbol{q}} D(\boldsymbol{p}^{\infty}(v)) = \frac{d}{2(1-2^{-d})} + \sum_{n=1}^{d} q_n \{E_n D(\boldsymbol{p}^{\infty}(v)) - n\}. \tag{5.12}$$

Lemma 5.1 with $\pi_k = p^{\infty}(k, v)$, implies that

$$\begin{aligned} \lim_{n\to\infty} E_n D(\boldsymbol{p}^{\infty}(v)) - n &= \sum_{k=1}^{\infty} U_k \\ &= 2 \sum_{1\leq k\leq j\leq\infty} \prod_{i=k}^{j} \frac{p^{\infty}(i,v)}{1-p^{\infty}(i,v)} \\ &= 2C(v/2). \end{aligned} \tag{5.13}$$

Equations (5.12) and (5.13) and the fact that $\sum_{n=0}^{j} q_n \to 0$ for any fixed j completes the proof of (3.27). The inequalities of (3.29) are proved as part of the lemma that follows. □

We close by presenting the asymptotics and a method for numerical calculation of

$$C(\mu) = \sum_{1\leq k\leq j\leq\infty} \prod_{i=k}^{j} a_i; \qquad 0 \leq \mu < 1$$

where

$$a_i = \frac{\mu^i}{1+i-\mu^i}.$$

Lemma 5.6: For $0 \leq \mu < 1$,

$$-\frac{\ln(1-\mu)}{\mu} - 1 \leq C(\mu) \leq -\frac{\ln(1-\mu)}{\mu} + 1.7. \tag{5.14}$$

Let

$$C_N(\mu) = \sum_{1\leq k\leq j\leq N} \prod_{i=k}^{j} a_i; \qquad 0 \leq \mu < 1. \tag{5.15}$$

For all $N \geq 1$ and $0 \leq \mu \leq 1$,

$$C_N(\mu) = \sum_{j=1}^{N} B_j \quad \text{where } B_1 = a_1, B_{j+1} = (1+B_j)a_j. \tag{5.16}$$

For $N \geq 2$,

$$0 \leq C(\mu) - C_N(\mu) \leq \frac{2}{N}\left(\frac{\mu^N}{1-\mu} + 1\right). \tag{5.17}$$

Proof: Dropping terms in the definition of $C(\mu)$ with $j > k$, and using the inequality $a_k \geq \frac{\mu^k}{1+k}$ and the Taylor series expansion of $\ln(1-\mu)$ yields that

$$C(\mu) \geq \sum_{k=1}^{\infty} a_k \geq \sum_{k=1}^{\infty} \frac{\mu^k}{1+k} = \frac{-\log(1-\mu)}{\mu} - 1. \tag{5.18}$$

On the other hand, we use the inequalities

$$a_k \leq \frac{\mu^k}{1+k} + \frac{1}{k} - \frac{1}{k+1} \quad \text{and} \quad \frac{a_{k+1}}{1-a_{k+2}} \leq \frac{1}{k},$$

the Taylor series expansion of $\ln(1-\mu)$, and the fact that a_k is monotone decreasing in k to obtain that

$$\begin{aligned} C(\mu) &= \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} a_k a_{k+1}(1 + a_{k+2} + a_{k+2}a_{k+3} + \cdots) \\ &\leq \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} a_k a_{k+1} \\ &\quad \cdot (1 + a_{k+2} + a_{k+2}^2 + a_{k+2}^3 + \cdots) \\ &\leq \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} \frac{a_k a_{k+1}}{1-a_{k+2}} \\ &\leq \sum_{k=1}^{\infty} \left(\frac{\mu^k}{1+k} + \frac{1}{k} - \frac{1}{k+1}\right) + \sum_{k=1}^{\infty} \frac{1}{k^2} \\ &= \frac{-\ln(1-\mu)}{\mu} + \frac{\pi^2}{6}. \end{aligned}$$

Equation (5.14) is established. To prove (5.17) use (5.15) and the fact that $a_i \leq 1/i$ for all i to get

$$\begin{aligned} 0 \leq C(\mu) - C_N(\mu) &= \sum_{j=N}^{\infty} \sum_{k=1}^{j} \prod_{i=k}^{j} a_i \\ &= \sum_{j=N}^{\infty} a_j \left(1 + a_{j-1} \sum_{k=1}^{j-1} \sum_{i=k}^{j-2} a_i\right) \\ &\leq \sum_{j=N}^{\infty} a_j(1 + (j-1)a_{j-1}) \leq 2 \sum_{j=N}^{\infty} a_j \\ &\leq 2 \sum_{j=N}^{\infty} \frac{\mu^j}{1+j} + \frac{1}{j} - \frac{1}{j+1} \\ &\leq \frac{2\mu^N}{(N+1)(1-\mu)} + \frac{2}{N}. \end{aligned}$$

Finally, it can be easily proved by induction on j that

$$B_j = \sum_{1 \leq k \leq j} \prod_{i=k}^{j} a_i$$

which implies (5.16). □

Acknowledgment

The authors are grateful to O. Schlunk for providing the first version of the simulation software.

References

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

[2] A. Borodin and J. E. Hopcroft, "Routing, merging, and sorting on parallel models of computation," *J. Comput. Syst. Sci.*, vol. 30, pp. 130–145, 1985. Also presented in part at the *14th Annu. ACM Symp. Theory Computing.*

[3] A. G. Greenberg and J. Goodman, "Sharp approximate models of adaptive routing in mesh networks," in *Teletraffic Analysis and Computer Performance Evaluation.* O. J. Boxma, J. W. Cohen, and H. C. Tijms, Eds. Elsevier, Amsterdam, 1986, pp. 255–270. Revised 1988.

[4] B. Hajek, "Bounds on evacuation time for deflection routing," *Distrib. Comput.*, vol. 5, pp. 1–6, 1991. Also, presented at the Conf. Syst. Sci., Johns Hopkins University, Baltimore, MD, Mar. 1989.

[5] M. T. Heath, Editor, *Hypercube Multiprocessors 1989,* Philadelphia, PA, 1987. Mathematical sciences section of the Oak Ridge National Laboratory and the SIAM activity group on supercomputing and linear algebra, *Siam Proc. Second Conf. Hypercube Multiprocessors;* Knoxville, TN, Sept. 29–Oct. 1, 1987.

[6] W. D. Hillis, *The Connection Machine.* Cambridge, MA: MIT Press, 1985.

[7] ——, "The connection machine," *Sci. Amer.*, vol. 246, no. 6, June 1987.

[8] A. Krishna, "Communication with few buffers: Analysis and design," Ph.D. thesis, Univ. Illinois at Urbana-Champaign, Dec. 1990. Also Tech. Rep. UIUC-ENG-90-2259.

[9] D. H. Lawrie and D. A. Padua, "Analysis of message switching with shuffle-exchanges in multiprocessors," in *Proceedings of the Workshop on Interconnection Network for Parallel and Distributed Processing,* pp. 116–123, 1980. Reprinted in IEEE Press book, *Interconnection Networks,* Wu and Feng. Eds. New York: IEEE Comp Soc. Press, 1984.

[10] N. F. Maxemchuk, "Regular mesh topologies in local and metropolitan area networks," *AT&T Tech. J.*, vol. 65, no. 7, pp. 1659–1685, Sept. 1985.

[11] ——, "Routing in the Manhattan street network," *IEEE Trans. Commun.*, vol. COM-35, no. 5, pp. 503–512, May 1987.

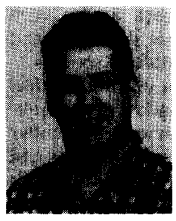
[12] B. Smith, "Architecture and applications of the HEP multiprocessor computer system," T. F. Tao, Ed., *Real-Time Signal Processing IV—Proc SPIE 298,* pp. 241–248. Society Photo-Optical Eng., 1981.

[13] T. Syzmanski, "An analysis of "hot-potato" routing in a fiber optic packet switched hypercube," in *Proc. IEEE INFOCOM '90,* vol. 2, San Francisco, CA, June 1990, pp. 918–926.

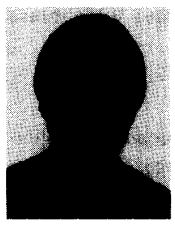
[14] X.-N Tan, K. C. Sevcik, and J.-W. Hong, "Optimal routing in the shuffle-exchange networks for multiprocessor systems," in *CompEuro 88—System Design: Concepts, Methods and Tools,* IEEE, Euromicro, April 1988, pp. 255–264. Published by IEEE Comput. Soc. Press, Washington, D.C.

[15] L. G. Valiant, "A scheme for fast parallel communication," *SIAM J. Comput.*, pp. 350–361, May 1982.

[16] L. G. Valiant and G. J. Brebner, "Universal schemes for parallel communication," in *Proc. 13th Annu. ACM Symp. Theory Comput.*, May 1981, pp. 263–277.



Albert G. Greenberg (M'88) received the Ph.D. degree in computer science, University of Washington, 1983, the B.A. degree in mathematics, Dartmouth College, 1978. He is a Member of Technical Staff, Mathematics Research Center, AT&T Bell Laboratories; joined in 1983. His interests include parallel processing, routing, communication networks, modeling and performance evaluation of computer systems.



Bruce Hajek (M'79–SM'84–F'89) received a B.S. in mathematics and an M.S. in electrical engineering from the University of Illinois in 1976 and 1977, and a Ph.D. in electrical engineering from the University of California at Berkeley in 1979. He is a Professor in the Department of Electrical and Computer Engineering and in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign, where he has been since 1979. He is the Editor-in-Chief (1989–1992), and was an Associate Editor, of the IEEE Transactions on Information Theory, and he served on the Board of Governors of the IEEE Information Theory Society, 1985–1990, 1992 to present. His research interest include communication and computer networks, stochastic systems, combinatorial and nonlinear optimization and information theory.