Performance of Congestion Control Mechanisms in Wormhole Routing Networks

CHRISTIAN ROCHE PRASASTH PALNATI MARIO GERLA
FABIO NERI EMILIO LEONARDI

UCLA Computer Science Dept.
3732H Boelter Hall
Los Angeles, CA 90095, USA.
email: roche@cs.ucla.edu

Abstract

In order to minimize latency in high-speed interconnection networks, the wormhole routing *technique can be employed. With this technique, a switch transmits an incoming message as soon as it receives it, without waiting for the entire message. The problem then is that a message stretches over several links and locks network resources, thus making a contention situation possible. Two principal congestion control mechanisms can be considered,* backpressure flow control *and* deflection routing. *The performance of these mechanisms depends both on the traffic characteristics and on the network topology. In order to compare them, we study analytically the behavior of a wormhole routing network model with random input traffic under both policies. We estimate the probability of collision between messages and express the average message transit delay as a function of the offered load. Simulation provides a good confirmation for the analytical results. This study gives us an understanding of the behavior of the system under different resource management policies.*

1 Introduction

An important concern for time-sensitive applications such as distributed computation over interconnected supercomputers or real-time data processing is the latency of the interconnection network. Since these applications require large amounts of data to be transmitted within short periods of time, the reaction time of the underlying communication system must be minimized. Connection-oriented techniques like ATM may not be best suited to this kind of application because the connection establishment and cell segmentation phases may introduce non-negligible delays in the transmission of the data. In order to minimize the latency of the network, the *wormhole routing* technique can be used [1, 2]. An example of a commercially available wormhole routing LAN is Myrinet [3]. Myrinet LANs use source routing and 8-bit wide data transmission to achieve data rates of up to 640 Mb/s. Their geographical reach is currently limited to a maximum link length of 25 meters. The ARPA-sponsored Supercomputer SuperNet (SSN) project at UCLA extends the reach of Myrinet LANs to campus area networks via an optical backbone network [4].

With wormhole routing, the data units are variable-length messages. A switch that receives the head of a message immediately forwards it if possible to the next switch on its path. It does not wait to receive the whole body of the message as with the classical *store-and-forward* policy. Thus, the latency associated with buffering the message is eliminated and the transit time of a message can be reduced to a minimum [4]. However, a message can be blocked by another message currently being transmitted on the desired output port.

Several approaches can be used to handle blocking. With *cut-through routing*, a buffer large enough to store a whole message is provided at each input port [5]. If a message cannot be transmitted it is buffered while waiting for the output port to become available. This reduces message loss but basically leads to the same type of congestion encountered in store-and-forward networks [2, 6].

A second approach consists in using a hop-by-hop *backpressure flow control* mechanism [7]. When the desired output port is unavailable, a message is blocked; a STOP signal is propagated upstream. This prevents the risk of buffer overflow but causes a lock up of network resources (i.e., buffers and switch ports). Under high load conditions, second and third order blockings and ultimately *deadlocks* occur.

A third approach is to use a *deflection routing* mechanism [7]. With such a mechanism, a message which finds the preferred output port busy is deflected to another port. Thus, messages may travel on longer paths and consequently suffer longer delays, but will not be lost due to buffer overflow. Most importantly, network resources are not locked up. An alternative to pure, asynchronous deflection routing is *delayed deflection*: a blocked message is temporarily buffered while waiting for its preferred output port to become available.

The performance of backpressure flow control depend on traffic characteristics, in particular message length. In contrast, the performance of deflection routing depend essentially on the network topology. Simulation results have been presented in [7]. Our goal in this paper is to evaluate and compare analytically the performance of these mechanisms.

This paper is organized as follow: in Section 2 we describe the different congestion control mechanisms. In Section 3 we present the network model and the modeling assumptions. We estimate the collision probability in Section 4 and assess the performance of the system under each policy in Section 5. Section 6 validates the theoretical results against simulation results and compare the respective performance of the mechanisms. Section 7 concludes the paper and indicates possible directions for future research.

2 Congestion control mechanisms

2.1 Backpressure flow control

The concept of backpressure flow control is illustrated in Figure 1. When a message is blocked, a *choke packet* is sent upstream to stop the transmission. A portion of the message is stored in a small *slack buffer* associated with the input port. When the link becomes free, a *release packet* is sent to resume transmission.

In order to be able to send the control packets, the backpressure flow control mechanism requires bidirectional links. Also, the switch must provide a small slack buffer to store the fraction of the

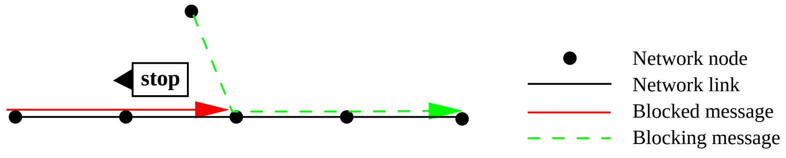

Figure 1: Backpressure flow control.

message that is “in flight”.

A critical problem of backpressure flow control is the occurrence of *deadlocks*. Several deadlock avoidance mechanisms have been proposed, including acyclic routing, timeouts and dynamic detection/resolution [8, 9]. In Myrinet, a type of acyclic routing called *Up/Down routing* is used [3].

2.2 Deflection routing

The concept of deflection routing is illustrated in Figure 2. When a message arriving at a switch finds its preferred output port busy, it doesn't wait. Rather, it is deflected from its shortest path and immediately transmitted on another available link. Thus, it takes a detour, which increases its path length and transit delay.

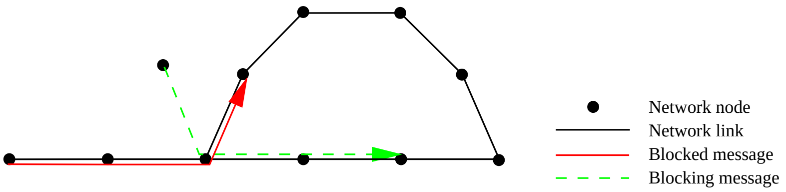

Figure 2: Deflection routing.

Deflection routing requires that there always exist an available output link. This can be achieved if the network has a *regular topology*, i.e., if each switch has as many inputs as outputs. Note that this condition is automatically met when the links are bidirectional. A drawback compared to backpressure flow control is that the message sequencing may not be preserved, since messages are randomly deflected [2, 7].

With deflection routing, deadlocks are not possible since messages are not blocked. However, *livelocks* can occur, i.e., situations where some messages are deflected indefinitely and never reach their destinations. As with deadlocks, such a situation must be detected and resolved. One possible solution is to limit the maximum number of hops that a packet can make. However, this poses the problem of choosing a suitable upper bound. As another approach, slotted networks use random choices of output ports to probabilistically break livelocks. In an asynchronous, unslotted case like ours, however, in overload conditions only one output port is free. Thus, random choices does not work. Some form of synchronization or message alignment may be desirable.

We consider the *delayed deflection* message alignment technique. When the head of a message arrives at a node and the preferred output port is unavailable, the node checks whether the message can be buffered in the input buffer until the preferred output port becomes available (i.e., until the remainder of the message using that port is transmitted). If the input buffer would overflow, then the message is deflected through another output port. Otherwise, the message is buffered and deflection is delayed. It is possible that by the time the output port becomes available, other messages waiting for the same output port may have arrived. One of these messages is then randomly chosen and the output port assigned to it. This message alignment mechanism introduces more choices of output port selection, thus resolving livelocks and improving the overall efficiency.

3 Network model

In order to evaluate the performance of the different congestion control mechanisms described above, we consider a simplified model of a wormhole routing network. The model is depicted in Figure 3.

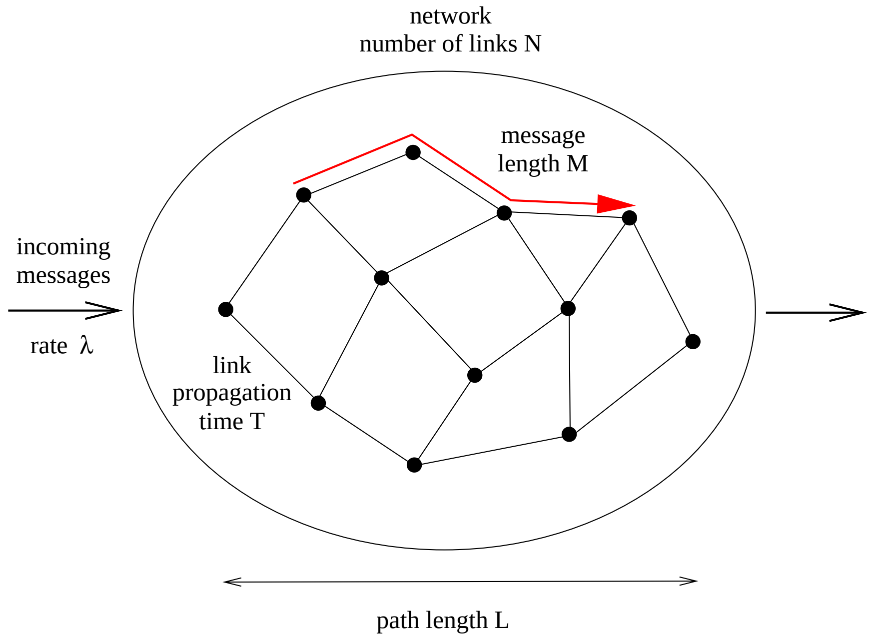

Figure 3: Wormhole routing network model.

The network we consider is composed of N unidirectional links with identical propagation times T and transmission capacities C. This model accommodates the case of bidirectional links by modeling each bidirectional link as two independent unidirectional links. There are H hosts communicating through the network. The messages are of fixed length M and cross the network along paths of length L, where L is the number of links. Time is expressed in link propagation time units. Likewise, M is expressed in number of links. More precisely, if b is the messages size in bits, then M is given by

$$M = \frac{b}{CT}. \tag{1}$$

In the deflection model, a key performance measure is the *average deflection cost* K. This quantity is strictly dependent on the topology and represents the average increase of the message path length introduced by a single deflection.

For the sake of simplicity, we make the following assumptions:

- We do not take into account the propagation delay of control packets in case of backpressure flow control, i.e., we assume that the messages are stopped instantaneously.

- We assume that the input traffic pattern is uniform and the network topology is such that the total load is evenly distributed throughout the system. All the links carry the same amount of traffic.
- We assume Poisson message arrival distribution with parameter λ.

In the following, we consider the system described above in a stationary regime under a given input load. We first derive the probability of collision and then the average message delay for various schemes.

4 Collision probability

We first introduce the dimensionless total input load z, defined as the ratio of the total incoming data rate to the nominal rate C of a link. This quantity can be viewed as the average number of hosts simultaneously transmitting. The theoretical maximum value of z (the saturation load) is H. From (1) we find

$$z = \frac{\lambda b}{C} = \lambda MT. \tag{2}$$

Next, we define m to be the average number of message headers in the system. From Little's result:

$$m = \lambda t, \tag{3}$$

where t is the average header transit delay (i.e., the time elapsed from the moment the first byte enters the network till the moment the first byte exits the network).

Recall that message length M was expressed in link segments. Thus, the average arrival rate expressed in link segments is $M\lambda$. Hence, the average number Λ of link segments in the system (i.e., the number of busy links) can be obtained by using Little's result again, provided that we know the average time that a link segment spends in the system.

In the case of pure deflection routing, each segment spends the same amount of time in the system, namely t. So the average number of busy links is simply

$$\Lambda_{\text{defl}} = M\lambda t = Mm. \tag{4}$$

This relation also holds for the case of delayed deflection as long as the probability of blocking remains small. This is the case when the size of the deflection buffer is small compared to the size of the messages.

Next, we consider the case of backpressure flow control. Note that once the head of a message has made its way through the system, the rest of the message crosses without blocking. In particular, the average transit time of each segment is a function of its position in the message, as shown on Figure 4.

In the simple case of constant message length M and path length L, the average transit time of a segment can be evaluated as $\xi t + (1-\xi)LT$, with

$$\xi = \begin{cases} 1 - \frac{M-1}{2L} & \text{if } M \leq L, \\ \frac{L+1}{2M} & \text{if } M \geq L. \end{cases}$$

So in the case of backpressure flow control we have

$$\Lambda_{\text{back}} = M\lambda\,(\xi t + (1-\xi)LT). \tag{5}$$

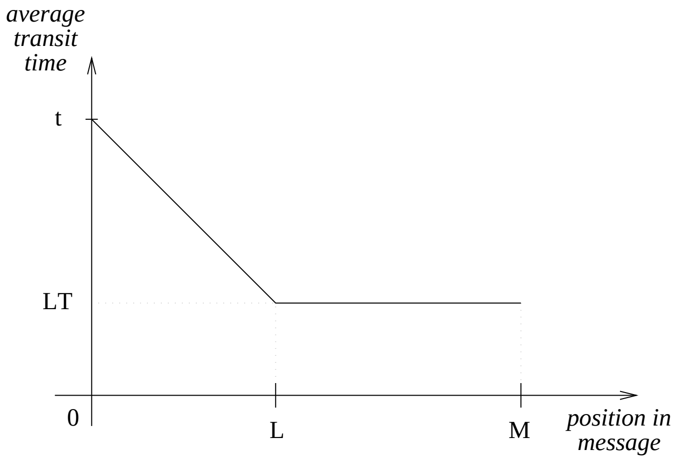

Figure 4: Transit time of each part of a message.

Next, we define the link collision probability p as the probability that the next link on a message path is busy. In order to derive p, we first assess the probability p_0 that any given link is busy. By definition, p_0 is the ratio of the average number of busy links to the total number of links, that is

$$p_0 = \frac{\Lambda}{N}. \tag{6}$$

The collision probability p depends on the time distribution of the message arrivals. In this study we consider the case of Poisson arrivals, which means that the probability of observing k arrivals in a time interval of length τ is

$$P_k = \frac{\omega^k}{k!}\, e^{-\omega},$$

where $\omega = \lambda\tau$. In order to evaluate p, we use for τ the characteristic time interval for message interaction. A message is present in the system (head to tail) for a time $t + MT$, so we set $\tau = t + MT$. Let us assume that during this time interval the message occupies an average of μ links, where $\mu \leq L$. Using Little's result, we can write

$$\Lambda = \mu\lambda\tau = \mu\omega.$$

We now consider a time interval of length τ on which there are more than one arrival and calculate the probability that a message be blocked by any message other than itself. Ignoring dependencies, p is given by the fraction of links occupied by other messages. Formally, this can be written

$$p = \frac{1}{N} \frac{\sum_{k=1}^{\infty} (k-1)\mu P_k}{\sum_{k=1}^{\infty} P_k}.$$

The numerator in this expression is the average number of link segments in the system belonging to messages other than the message in question. We finally obtain the following expression for p:

$$p = \frac{\mu}{N}\left(\frac{\omega}{1-e^{-\omega}} - 1\right) = p_0\left(\frac{1}{1-e^{-\omega}} - \frac{1}{\omega}\right). \tag{7}$$

5 Analysis of the mechanisms

5.1 Deflection routing

Let us first consider the case of pure deflection routing. We denote by l the average path length of the messages expressed in number of links. The minimum path length of a message is L and is increased by K each time the message experiences a deflection.

The collision probability is p, which means that on a path of length l a message experiences on average lp deflections. Hence, we have the relation

$$l = L + lpK,$$

which yields

$$l = \frac{L}{1 - pK}.$$

Since in the case of pure deflection routing a message is never blocked, its transit time is directly proportional to the average number of links it crosses, that is

$$t = \frac{LT}{1 - pK}. \tag{8}$$

Equations (7) and (8) form a system of two equations in the two unknown p and t, with z defined in (2) playing the role of a parameter. This system is non-linear and cannot be solved in closed form. However, we can use numerical techniques to solve it for different values of the load z, as shown in Section 6.2. Also, we can use approximations to linearize the system for small values of z, as shown in Section 6.3.

5.2 Backpressure flow control

In the case of backpressure flow control, the average path length is fixed and is equal to L. So, a message experiences an average of Lp collisions. In order to evaluate the waiting time upon blocking, we define the *average progress rate* (in links per sec) of a message as

$$v = \frac{L}{\xi t + (1 - \xi)LT}.$$

On average, a message is blocked in the middle and must therefore wait for the blocking message to travel $M/2$ links. Thus, the average waiting time upon blocking can be written as

$$w = \frac{M}{2} \frac{\xi t + (1 - \xi)LT}{L}.$$

Hence, we can write

$$t = LT + Lp \frac{M}{2} \frac{\xi t + (1 - \xi)LT}{L},$$

which gives

$$t = LT \frac{1 + (1 - \xi)pM/2}{1 - \xi pM/2}. \tag{9}$$

5.3 Delayed deflection

In the case of delayed deflection, each node has on each input port a buffer of size B, with $0 < B < M$. When a message experiences a collision, its time to wait is estimated and the message is deflected only if the wait exceeds the available buffer. So if we denote $\gamma = B/M$, we have upon each collision:

$$\begin{array}{lcll} \text{With probability} & \gamma & \rightarrow & \text{buffering,} \\ \text{With probability} & 1-\gamma & \rightarrow & \text{deflection.} \end{array}$$

In order to evaluate the average path length of the messages, we can use the result of Section 5.1 with a reduced probability of deflection $p' = (1-\gamma)p$. Thus, we obtain

$$l = \frac{L}{1-(1-\gamma)pK}.$$

Now when a message is blocked, its average waiting time is $BT/2$, so we can write

$$t = lT + l\gamma p\,\frac{BT}{2}.$$

Finally, we obtain

$$t = LT\,\frac{1+\gamma^2 pM/2}{1-(1-\gamma)pK}. \tag{10}$$

6 Numerical results

In order to verify the validity of the analytical model, we compare in this section the analytical results against simulation.

6.1 Simulation model

For the simulation runs we use a 64-node bidirectional network with *torus* and *shuffle* topologies [2]. The connectivity of the network with both topologies is 4, which means that there are four bidirectional transit ports per node. The corresponding number of links in the analytic model is $N = 4 \times 64 = 256$. The dimension parameters of the simulation model are gathered in Table 1.

The average deflection cost K is defined as the increase in path length that a message experiences when it is deflected from its shortest path. This value is averaged over all source/destination pairs and is weighed by link traffic. The computation of K can be made exactly for a given topology with a given input traffic pattern. It does not depend on the assumptions made in this study.

The network uses shortest-path routing, i.e., a message is always transmitted when possible on a link that brings it closer to its destination. If several links are possible, one is chosen at random. Each node sends with the same average rate to all the other nodes.

It is interesting to note that with this uniform traffic pattern, the total load is equally distributed on all the links in the case of the torus topology but not in the case of the shuffle topology. In this latter case, there is a ratio of 2 to 3 between the traffic rates at the least and most loaded links. Also, even though we carried the analysis considering paths of fixed length L, the simulation runs

Symbol	*Quantity*	*Shuffle*	*Torus*
T	Propagation time on a link	1	
H	Number of hosts	64	
N	Number of links	256	
M	Message length	12	
B	Deflection buffer size	1.2	
L	Average path length	3.429	4.063
K	Average deflection cost	1.385	0.942

Table 1: Dimensions of the simulation model.

are conducted with paths of different lengths. Now L represents the average path length under the considered traffic pattern.

We will see below that, in spite of these differences between the analytic and simulation models, the simulation results still provide a very satisfying match with the analysis. This confirms the validity of the analytic model when these simplifying assumptions are relaxed.

6.2 Simulation results

Figures 5 shows the curves obtained by simulation with the above system configuration in the case of backpressure flow control and pure deflection routing, and compares them with the theoretical results obtained using numerical techniques. These curves depict the average transit time t of the messages as a function of the total input load z. We see that the analytical model provides quite a satisfying approximation of the system performance.

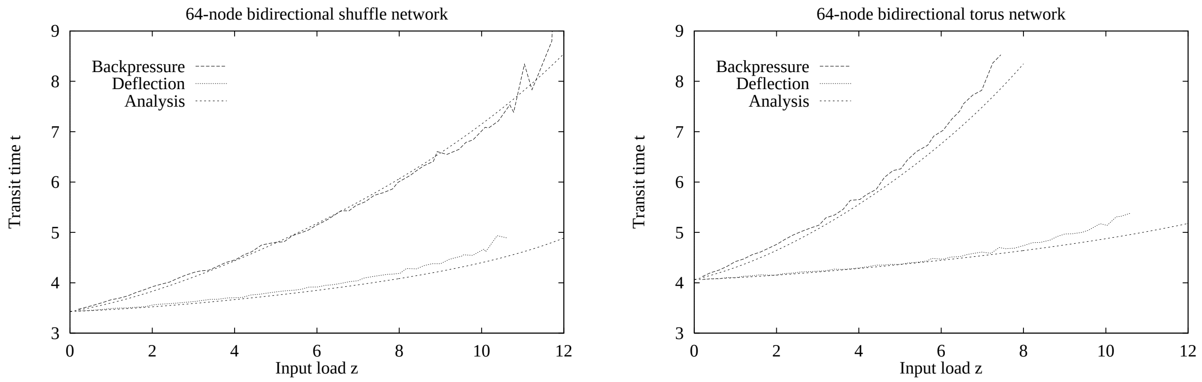

Figure 5: Backpressure flow control *vs* pure deflection.

The deviation that can be observed at higher loads (above 15% of capacity) comes from the fact that when the system's load increases, deadlocks and livelocks occur more often. Then the corrective measures that must be undertaken to resolve these situations (essentially destroying the corresponding messages) steer the real system's behavior away from the ideal behavior described by the analytical model. When the offered load exceeds 20% of capacity, the network becomes unstable and its performance collapses, due to the number of deadlocks or livelocks that the system suffers. This shows that the simplest operation mode, with unrestricted routing and with no input control, is not applicable in a real situation. This issue will be further discussed in the next section.

Figure 6 compares the pure deflection and delayed deflection mechanisms with $\gamma = 10\%$. We see that, although we cannot distinguish quantitatively between the two mechanisms, the analytical model still captures the general shape and respective position of the curves.

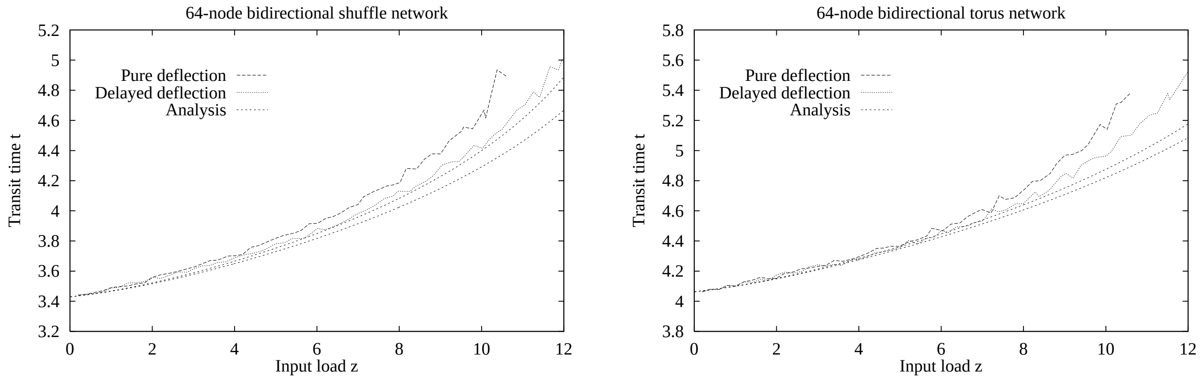

Figure 6: Pure deflection *vs* delayed deflection.

6.3 Performance comparison

In order to be able to compare the performance of the mechanisms, we need to define a simple performance criterion. One possible choice is the slope of the curve $t(z)$ at $z = 0$. This quantity captures the respective position of each curves and thus the relative performance of each mechanism, at least at low load. To evaluate this quantity we approximate the expression of the collision probability given by equation (7) to the first order, hence

$$p \approx \frac{p_0}{2}.$$

Then t appears in each case as the solution of a second-degree polynomial in z and can be expressed in a closed form. To compute the slope of this polynomial at the origin is then straightforward. The resulting quantity is the exact expression of the desired performance criterion. In the case of pure deflection routing, we have

$$\left.\frac{dt}{dz}\right|_{z=0} = \frac{TL^2K}{2N}.$$

In the case of backpressure flow control, we obtain

$$\left.\frac{dt}{dz}\right|_{z=0} = \frac{TL^2M}{4N}.$$

We can see that these two expressions are similar, with K and $M/2$ playing an equivalent role. This shows that the relative performance of backpressure flow control compared to pure deflection routing depends essentially on the average message length compared to the average deflection cost. If $M/2$ is greater than K, then deflection routing will exhibit better performance. On the other hand, if it is smaller, the backpressure flow control will exhibit better performance.

6.4 Extensions of the model

The analysis so far was conducted using a simplified wormhole routing network model. The maximum offered load was limited due to deadlock and livelock. In [7], a more detailed model of a

Myrinet network is studied through simulation. It includes deadlock avoidance mechanisms such as Up/Down routing and virtual channels and considers variable-size messages. The interested reader will find a detailed description of the simulator along with an extensive presentation of results in [7]. In this paper, we limit ourselves to reporting the transit delay results for the 64-node bidirectional torus network (see Figure 7).

In this figure, the x axis indicates the average host load (i.e., z/H). The y axis indicates the average message transit delay, measured from the time the first byte enters the network to the time the last byte is delivered to the destination (i.e., $t + MT$). This quantity is measured in byte transmission times. The considered 1 km links at 640 Mb/s represent 400 byte times (i.e., $T = 400$). Several policies are considered, namely deadlock-free virtual channels (VC), deflection routing with different deflection buffer sizes (DF) and Up/Down deadlock-free restricted routing (UP). The message length is geometrically distributed with an average of 5000 bytes (i.e., $M = 12.5$) and a upper limit of 10000 bytes.

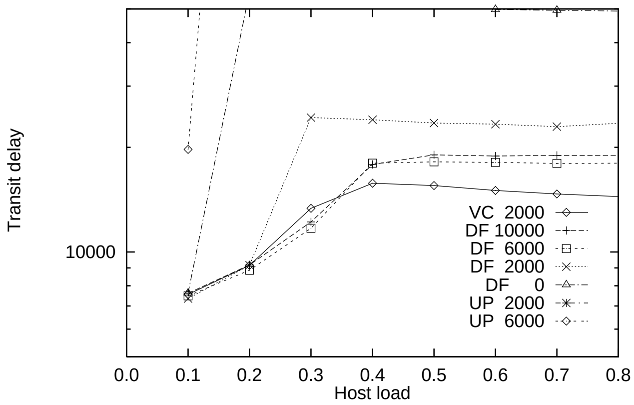

Figure 7: Average message delay *vs* host load.

These curves can be compared to those in Figures 5 and 6 in the case of torus topology. We note that pure deflection delay for load .1 (7,000 byte times) is consistent with the value in Figure 5 (namely, $(4.5 + 12) \times 400 = 6,600$). The backpressure delay with Up/Down routing (for deadlock prevention) is somewhat higher due to the longer paths and link loads compared to shortest path routing.

The most interesting information provided by Figure 7 is that higher loads can be achieved by employing suitable deadlock and livelock avoidance mechanisms. Therefore, the analytical models developed in this paper can be profitably extended to incorporate the behavior of more complex flow control strategies. This extension is considered as a direction for future research.

7 Conclusion

Wormhole routing is a promising technique for low-latency, high-speed interconnection networks. However, this technique introduces challenging problems including resource contention and deadlock occurrence. Two principal mechanisms can be employed to alleviate congestion in wormhole routing networks, namely backpressure flow control and deflection routing.

In this paper, we analyzed and compared the performance of these mechanisms in a general network configuration. By estimating the probability of collision in the system for a given input load, we are able to express the average message transit time as a function of the offered load. Simulation results show a good match with the analytical results.

This study allows us to compare the respective performance of the mechanisms and decide which one is best suited to a given network configuration. It also helps understanding the influence of the system dimensions on its overall performance and thus predict how the system will scale.

Several research directions can be considered from here. One direction consists in refining the model and relaxing some assumptions in order to get a closer match to a real system. Another interesting issue would be to apply the analytical techniques developed here to other networking techniques and compare them with one another. Also, work is currently in progress in developing the analytical model in order to capture the behavior of different deadlock/livelock avoidance mechanisms.

References

[1] L. Li and P. McKinley, "A survey of wormhole routing techniques in direct networks," *IEEE Computer*, vol. 26, no. 2, pp. 62–76, Feb. 1993.

[2] C. Baransel, W. Dobosiewicz and P. Gburzynski, "Routing in multihop packet switching networks: the Gb/s challenge," *IEEE Network Magazine*, vol. 9, no. 3, 1995.

[3] N. Boden, D. Cohen *et al.*, "Myrinet: a gigabit-per-second local-area network," *IEEE Micro*, vol. 15, no. 1, Feb. 1995.

[4] L. Kleinrock, M. Gerla *et al.*, "The Supercomputer Supernet testbed: a WDM based supercomputer interconnect," *IEEE JSAC*, vol. 14, no. 6, June 1996.

[5] P. Kermani and L. Kleinrock, "Virtual cut-through: a new computer communication switching technique," *Computer Networks*, vol. 3, no. 3, pp. 267–286, Sept. 1979.

[6] R. Jain, "Congestion control in computer networks: issues and trends," *IEEE Network Magazine*, vol. 4, no. 3, pp. 24–30, May 1990.

[7] E. Leonardi, F. Neri, M. Gerla *et al.*, "Congestion control in asynchronous, high-speed wormhole routing networks," *IEEE Communication Magazine*, Nov. 1996.

[8] W. Dally and C. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," *IEEE Transactions on Computers*, vol. C-36, no. 5, pp. 547–553, May 1987.

[9] P. Palnati, M. Gerla and E. Leonardi, "Deadlock-free routing in an optical interconnect for high-speed wormhole routing networks," in *Proceedings of 1996 International Conference on Parallel and Distributed Systems*, Tokyo, June 1996.