On the Performance of Routing Algorithms in Wormhole-Switched Multicomputer Networks

A. Shahrabi
School of Computing and Mathematical Sciences
Glasgow Caledonian University
Glasgow G4 0BA, U.K.
a.shahrabi@gcal.ac.uk

M. Ould-Khaoua
Department of Computing Science
University of Glasgow
Glasgow G12 8RZ
U.K.
Mohamed@dcs.gla.ac.uk

Abstract

This paper presents a comparative performance study of adaptive and deterministic routing algorithms in wormhole-switched hypercubes and investigates the performance vicissitudes of these routing schemes under a variety of network operating conditions. Despite the previously reported results, our results show that the adaptive routing does not consistently outperform the deterministic routing even for high dimensional networks. In fact, it appears that the superiority of adaptive routing is highly dependent to the broadcast traffic rate generated at each node and it begins to deteriorate by growing the broadcast rate of generated message.

1. Introduction

A number of performance studies [3], [14], [15] in wormhole-routed networks have shown that adaptive routing can not only achieve a higher throughput compared to deterministic routing, but also a lower latency. The performance advantages of adaptivity are pronounced even in high-dimensional networks such as hypercubes. However, these studies have been carried out in the context of unicast (or point-to-point) communication only. Many real-world parallel applications generate broadcast workloads which have a significant impact on network performance [9], [12]. It is therefore critical to consider this type of traffic in any performance study in order to obtain a more realistic picture of the important factors that affect network performance.

Most network performance evaluation studies have been conducted by means of software simulation [2], [4], [11]. Studying the relative performance merits of routing algorithms in the presence of broadcast traffic using simulation techniques is, however, limited by the excessive computation times required to run large simulations. Analytical modelling, in contrast, offers a cost-effective and versatile tool to carry out such a study typically requiring a far lower computational load. This study uses the analytical models recently proposed in the literature [18], [19] to present the first comparative performance evaluation of adaptive and deterministic routing algorithms in hypercubes under different traffic conditions which include a mixture of broadcast and unicast communication components.

The rest of this paper is organised as follows. Section 2 briefly gives some preliminaries to this study. Section 3 presents the results and compares the performance of adaptive against deterministic routing under various working conditions. Finally, section 4 concludes this paper.

2. Preliminaries

2.1. Routing algorithms

Routing algorithms establish the path followed by each message. They are broadly classified as *deterministic* and *adaptive* algorithms [9], [11]. In deterministic routing, the source and destination addresses of each arriving packet deterministically select an output channel. Adaptive routing, on the other hand, exploits the fact that there might be more than one path between any source and destination node in a multi-dimensional network. The decision as to which output channel should be selected for a packet is based on dynamic factors such as current network traffic, channel status, and the distance from the destination node.

The simplest deterministic routing algorithm consists of reducing an offset to zero before considering the offset in the next dimension. This routing algorithm is known as *dimension-order* algorithm. The dimension-order routing algorithm routes packet by crossing dimensions in strictly increase (or decrease) order, reducing to zero the offset

Proceedings of the 2005 11th International Conference on Parallel and Distributed Systems (ICPADS'05)
0-7695-2281-5/05 $20.00 © 2005 IEEE

in one dimension before routing in the next one. Dimension-order routing is very popular and is known *e*-cube routing for hypercubes [9].

Among all adaptive routing algorithms have been discussed in literature [4], [5], [8], the Duato algorithm is perhaps the most attractive since it requires a limited number of virtual channels to ensure deadlock freedom. It has therefore been widely studied and is accepted as an approach to adaptive routing with minimal resource requirements. The Cray T3E [17], and the reliable router [6] are two examples of recent practical systems that have adopted Duato routing algorithm.

2.2. Broadcast algorithm

Our present study focuses on the *Double Tree* (DT) broadcast algorithm, proposed by McKinley and Trefftz [13], for multiple-port wormhole-routed hypercubes. The main advantage of the DT algorithm stems from the fact that DT algorithm divides the hypercube into two parts, and builds in each part a "reduced height" tree. In the DT algorithm, a source node, say A, initiates a broadcast operation by sending a copy of the broadcast message, referred to here as the *diagonal* message, to node $\bar{A}$ whose address is the "bit-wise" complement of the node A's address. In the next step, nodes A and $\bar{A}$ become the roots of partial spanning binomial trees, along which copies of the broadcast message propagate to all the other network nodes. The DT algorithm broadcasts a message to all the nodes in $\lceil n/2 \rceil$ routing steps. The partial spanning tree rooted at A and $\bar{A}$ are called the *forward* and *backward* tree, respectively [13].

3. Results and discussions

3.1. Network traffic

The traffic distribution exhibited by parallel applications is an important factor that strongly affects network performance. Unicast communication involves only two nodes: the source and destination. The uniform traffic pattern is a typical example of unicast communication, which has been widely considered when analysing network performance [1], [2], [10]. Broadcast communication is often needed in scientific computations to distribute large data arrays over system nodes in order to perform various data manipulation operations. From a system perspective, on the other hand, broadcast communications are required in control operations such as global synchronisation. Furthermore, in the distributed shared-memory (DSM) paradigm, broadcast communication is used to support shared data invalidation and updating procedures required for cache coherence protocols [7]. The importance of broadcast communication is evidenced by its inclusion in the Message passing Interface (MPI) [16].

The question that is of key importance at this point is just how likely it is that the generated traffic at each node is due to broadcast operations. To provide an answer, it would be necessary to use the representative traces from intended applications. However, it is very difficult to fix a typical ratio of broadcast to unicast messages as this is highly dependent on the underlying system and the communication requirements of parallel applications; this varies not only from one network to another but also among different applications. For instance, the broadcast traffic rate in the DSM model is likely to be much higher than that in a message-passing system [9], because the former uses broadcast communication heavily to support shared data invalidation and the updating procedures required for cache coherence protocols.

A broadcast message is delivered to every node in the network using the broadcast algorithm described in Section 2.2. A unicast message is sent to other nodes in the network with equal probability. When a message is generated in a given source node, it has a finite probability of being a broadcast message. In the analytical models developed in [18] and [19], the broadcast traffic rate at each source node is denoted by β. So, a message has the probability $(1-\beta)$ to be a unicast message.

As mentioned above, determining realistic values for β requires the use of representative traces from intended applications but, in any case, existing literature tends to use a figure in the range of 1-10% of generated messages [9]. Although this may seem modest, small values of β correspond to a large, possibly even dominant, component of broadcast traffic when the network is large. Yet, despite its importance, this crucial factor has received little attention in the performance evaluation studies reported in the literature.

Based upon the equations provided for channel traffic in analytical models presented in [18] and [19], Figures 1 illustrates how broadcast traffic is a dominating component of traffic on network channels. It compares the amount of broadcast and unicast traffic on each network channel in a hypercube of dimension 12 for two different values of β = 0.01 and 0.07, with V = 4 virtual channels per physical channel and M = 64 flits message length under DT broadcast traffic. While the broadcast component of traffic on a given network

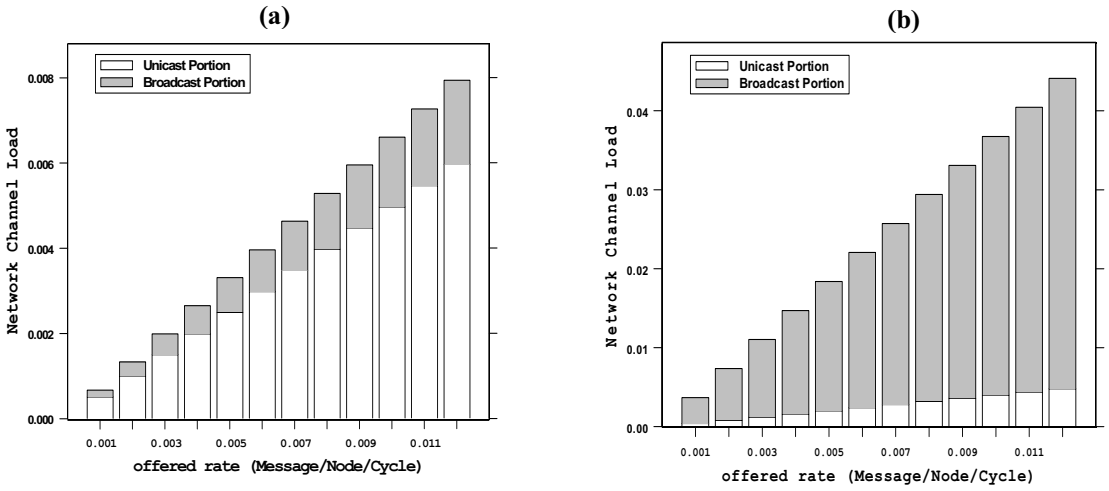


Figure 1: Traffic load on each channel in a hypercube of dimension 10 with 4 virtual channels, 64 flits message length and a) 0.01, b) 0.2 broadcast traffic rate.

channel is almost equal to the unicast load when β = 0.01 (Figure 1(a)), it is the dominant component when the broadcast traffic rate, β, increases to 0.07 (Figure 1(b)).

3.2. Effects of broadcast traffic

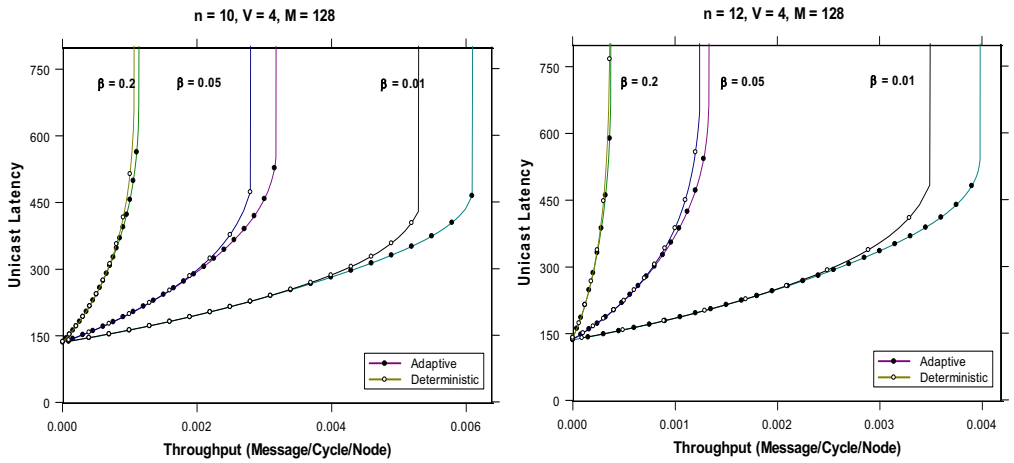
Figure 2(a) compares the average unicast latency of adaptive and deterministic routing in a 10-dimensional hypercube as a function of network load in the presence of DT broadcast traffic; considering three different rates, β = 0.01, β = 0.05, and β = 0.2. The network is assumed to have V = 4 virtual channels per physical channel and message length is M = 64 flits. Figure 2(b) shows the same results for a 12-dimensional hypercube with the same number of virtual channels per physical channel and M = 64 flits message length.

These figure indicate a performance advantage for adaptive routing when the broadcast traffic rate is very low (e.g. β = 0.01). In this situation, both adaptive and deterministic routing have, as expected, the same message latency under light and moderate traffic (see Figure 2(b), β = 0.01). When the offered traffic rises above 70% of maximum sustained throughput however, adaptive routing not only has a lower latency but also achieves higher throughput. In Figure 2(b), for example, the maximum sustained throughput of adaptive routing is 11% higher than that of deterministic routing.

Increasing the broadcast traffic rate changes these results and brings the performance of the hypercube with adaptive routing gradually closer to that with deterministic routing, offsetting any advantage of having additional flexibility to determine a path between the current node and destination. For instance, in Figure 2(b), for β=0.05 the ratio of deterministic to adaptive throughput is 0.93 whereas it increases to 0.98 when increasing β to 0.2. At β=0.2, the performance of a 12-dimensional hypercube with adaptive routing is practically the same as with deterministic routing.

Figure 3(a) and 3(b) show respectively the maximum sustained throughput for a 10-dimensional and a 12-dimensional hypercube under unicast and DT broadcast traffic. In both cases message length is 64 flits, and the number of virtual channels per physical channel is set first to V = 4 and then to V = 8. These graphs show that adaptive and deterministic routing achieve comparable maximum sustained throughput as the broadcast traffic rate increases.

There are a number of possible explanations. Firstly, as mentioned previously, although the values of broadcast traffic rate, β, are relatively small, it is important to remember that this value is the broadcast rate of each node; small values of β correspond to large amounts of broadcast traffic when the network



(a) **(b)**

Figure 2: The effect of different broadcast traffic rate on the performance of adaptive and deterministic routing in a) 10-dimensional hypercube with 4 virtual channels and 128 flits message length and b) 12-dimensional hypercube with 4 virtual channels and 64 flits message length.

size is large. As a consequence, when the output channels of a router in the network are flooded with broadcast messages the degree of adaptivity (i.e., the number of available alternatives paths) decreases, and as a result adaptive routing exhibits comparable performance behaviour to that of deterministic routing. Secondly, in broadcast algorithms based on the unicast-based approach, no matter which broadcast algorithm, switching method or routing scheme is employed, a huge amount of generated traffic on network channels is due to those broadcast and replicated messages which are only one hop away from their destinations: referred to as "one-step" messages. Obviously, adaptivity is meaningless for one-step messages as there is only one output channel that this type of message can take to reach its destination.

4. Conclusions

This paper has presented a comparative performance study of adaptive and deterministic routing in wormhole-routed hypercubes in the presence of broadcast traffic and investigated the performance of these routing schemes under a variety of network operating conditions. It has also been shown that the performance advantage of adaptive routing is highly dependent on the amount of broadcast traffic present in the network. Previous performance results, reported in the context of unicast communication, have shown superiority of adaptive over deterministic routing, especially in high-dimensional networks such as hypercubes. However, this paper has demonstrated that adaptive routing does not consistently outperform deterministic routing even for high dimensional networks, when broadcast traffic is taken into consideration. Deterministic routing is able to achieve latency and throughput comparable to those achieved by adaptive routing even for relatively small values of broadcast traffic rate. These results show that adaptivity does not always improve network performance and its relative superiority reduces as the broadcast traffic rate increases.

References

[1] S. Abraham, "Interconnection networks dimensions in design," Proceedings Workshop of International Conference on Parallel Processing, pp. 45-51, 1996.

[2] V. Adve, M.K. Vernon, "Performance analysis of mesh interconnection networks with deterministic routing," IEEE Transactions on Parallel and Distributed Systems, Vol. 5, No. 3, 1994, pp. 225-246.

[3] R. Boppana, and S. Chalasani, "A framework for designing deadlock-free wormhole routing algorithms," IEEE Transaction on Parallel and distributed Systems, Vol. 7, No. 2, February 1996, pp. 169-183.

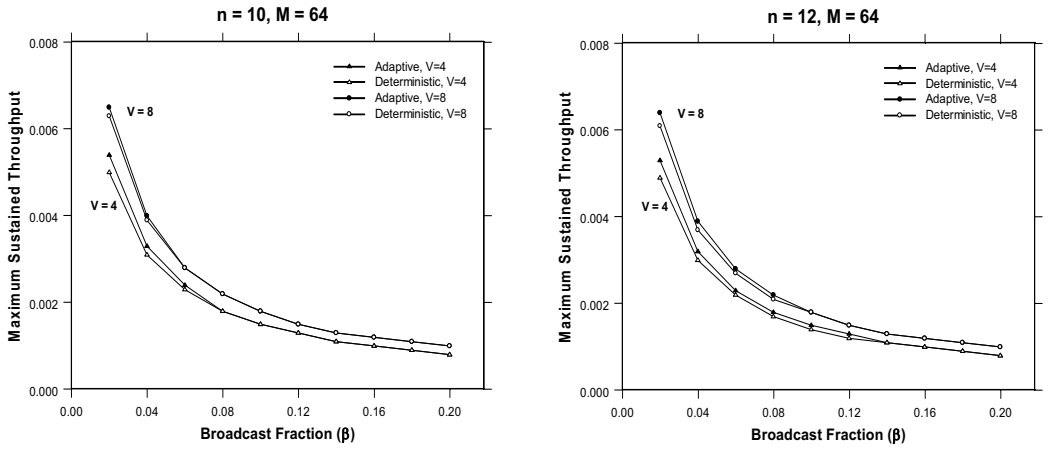


Figure 3: The effect of broadcast traffic rate (β) on the maximum sustained throughput for different numbers of virtual channels per physical channel (V = 4 and 8) in a a) 10 and b) 12-dimensional hypercube with 64 flits message length.

[4] Y. M. Boura, and C. R. Das, "Efficient fully adaptive routing in n-dimensional mesh," Proceedings of the 14th International Conference on Distributed Computing Systems, 1994.

[5] A. A. Chien, and J. H. Kim, "Planar Adaptive Routing: Low-cost Adaptive Networks for Multiprocessors," Proceedings of International Symposium on Computer Architecture, 1992, pp. 268-277.

[6] W. J. Dally, R. Dennison, D. Harris, K. Kan, and T. Xanthopoulos, "The reliable router: a reliable and high-performance communication substrate for parallel computers," Proceedings of the Workshop on parallel Computer Routing and Communication, May 1994, pp. 241-255.

[7] W. Dally, and B. Towles, "Principles and practices of Interconnection Networks," Morgan Kaufmann, 2004.

[8] J. Duato, "A Necessary and Sufficient Condition for Deadlock-Free Adaptive Routing in wormhole Networks," IEEE Transaction on Parallel and Distributed Systems, Vol. 6, No. 10, October 1995, pp. 1055-1067.

[9] J. Duato, S. Yalamanchili, L. Ni, Interconnection networks: An engineering approach, IEEE Computer Society Press, 1997.

[10] A. Khonsari, A. Shahrabi, and M. Ould-Khaoua, "A performance model of a true fully adaptive routing algorithm", Proceedings of 10th ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS' 02), Texas, October 2002.

[11] M.P. Malumbres, J. Duato, and J. Torrellas, An efficient implementation of tree-based multicast routing for distributed shared memory multiprocessor, Proc. 8th IEEE Int. Symp. Parallel & Distributed Processing, 1996.

[12] P. K. McKinley, Y. Tsai, and D. Robinson, "A Survey of Collective Communication in Wormhole-Routed Massively Parallel Computers," Technical Report MSU-CPS-94-35, Communications Research Group, Michigan State University, June 1994.

[13] P. K. McKinley, and C. Trefftz, Efficient broadcast in all-port wormhole-routed hypercubes, Proc. Int'l Conf. on Parallel Processing, 1993, pp. 288-291.

[14] D. Miller, and W. A. Najjar, "Preliminary evaluation of a hybrid deterministic/adaptive router," Proceedings of Parallel and Computing, Routing and Communication Workshop, June 1997.

[15] D. Miller, and W. A. Najjar, "Empirical evaluation of deterministic and adaptive routing with constant-area routers," Proceedings of Conference on Parallel Architecture and Compilation Techniques (PACT), November 1997.

[16] M. P. I. Forum, "MPI: A message-passing interface standard," March 1994.

[17] S. L. Scott, and G. M. Thorson, "The Cray T3E network: adaptive routing in a high performance 3D torus," Proceedings of the Symposium on High Performance Interconnects (Hot Interconnects 4), August 1996, pp. 147-156.

[18] A. Shahrabi, M. Ould-Khoua and L. Mackenzie "Latency of Double-Tree Broadcast in Wormhole-Routed Hypercubes," Proceedings of International Conference on Parallel Processing (ICPP'01) IEEE Computer Society, Valencia, Spain, September 3-7, 2001, pp. 401-408.

[19] A. Shahrabi, M. Ould-Khoua, L. Mackenzie, "An Analytical Model of Wormhole-Routed Hypercubes under Broadcast Traffic ", Performance Evaluation, Vol. 53, No. 1, June 2003, pp. 23-42.