Towards A Scalable Broadcast in Wormhole-Switched Mesh Networks

A. Y. Al-Dubai, M. Ould-Khaoua and L.M. Mackenzie Department of Computing Science University of Glasgow Glasgow G12 8RZ, U.K. Email: {aldubai, mohamed, lewis}@dcs.gla.ac.uk

ABSTRACT

Broadcast algorithms for wormhole-switched meshes have been widely reported in the literature. However, most of these algorithms handle broadcast in a sequential manner and do not scale well with the network size. As a consequence, many parallel applications cannot be efficiently supported using existing algorithms. Motivated by these observations, this paper presents a new broadcast algorithm based on our previously proposed Coded Path Routing (or CPR for short) [1]. The main feature of the proposed algorithm lies in its ability to perform broadcast operations with a high degree of parallelism. Furthermore, its performance is insensitive to the network size, i.e., only two message-passing steps are required to implement a broadcast operation irrespective of the network size. Results from a comparative analysis reveal that the new algorithm exhibits superior performance characteristics over those of the well-known Recursive Doubling, Extending Dominating Node and Network Partitioning algorithms.

Keywords

Interconnection Networks, Routing, Wormhole Switching, Collective Communication, Communication Latency.

1. INTRODUCTION

Maximising the performance of parallel systems requires matching message-passing algorithms and application characteristics with a suitable underling *interconnection network*. The mesh has been one of the most popular interconnection networks for multicomputers due to its desirable properties, such as ease of implementation, recursive structure, and ability to exploit communication locality found in many parallel application to reduce message latency.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires

prior specific permission and/or a fee. SAC 2002, Madrid, Spain. © 2002 ACM 1~58113-445~2/02/03...\$5.00. The M-machine, J-Machine, and iWarp are examples of practical systems that are based on the mesh topology.

Collective communication such as broadcast which refers to the delivery of the same message originating from a given source to all network nodes, is important in many real-world parallel applications found in the areas of Science and Engineering [2, 11]. For instance, broadcast communication is often needed in scientific computations to distribute large data arrays over system nodes in order, for example, to perform various data manipulation operations. Furthermore, it is required in control operations such as global synchronisation and to signal changes in network conditions, e.g., faults. In the distributed shared-memory paradigm, broadcast communication is often used to support shared data invalidation and updating procedures required for cache coherence protocols [3].

Several broadcast algorithms have been proposed in the literature for the wormhole-switched mesh [2, 7, 8, 9, 10]. These algorithms try to minimise the broadcast latency by reducing the number of message-passing steps, i.e. the number of exchanges, required to perform a broadcast operation. However, most of these algorithms do not scale well with the system size as they suffer from the degrading effects of the start-up latency, the required time to handle a broadcast message at both the source and destination nodes [3], especially when the network size is large. This is because the number of message-passing steps that is required to complete a broadcast operation usually depends on the network size. In the message-passing step, the communication latency is dominated by the start-up latency [7]. Therefore, scalable and highly parallel broadcast algorithms should have minimum message-passing steps to minimise the degrading effects of the start-up latency on network performance.

This paper uses the Coded Path Routing (or CPR for short) that has been proposed in [1] as a new approach for designing efficient broadcast algorithms for the mesh. A unique feature of the CPR is that a message with a single address can be delivered to an arbitrary number of destinations with single start-up latency only. Specifically, the CPR is used to devise a new broadcast algorithm for the multiple-port 2-dimensional mesh. Owing to the properties of the CPR, the proposed algorithm requires only two message-passing steps to implement a broadcast operation, irrespective of the system size. Results from analytical and simulation studies presented below reveal that the new algorithm exhibits superior performance characteristics over the well-known Recursive Doubling, Extending Dominating Node and Network Partitioning algorithms proposed in [2], [9] and [8], respectively. Most existing algorithms [2, 8, 9, 10, 11] require a number of message-passing steps that is proportional to the network size and implement broadcast operations in a highly sequential manner, i.e. most nodes in the network receive the broadcast message sequentially. There has been hardly any attempt to provide broadcast algorithms with a high degree of scalability and parallelism. In an effort to address these shortcomings, this paper presents an efficient broadcast algorithm, which overcomes the limitations of the existing algorithms. The remainder of this paper is organised as follows. Section 2 introduces the CPR. Section 3 is devoted to the system model. Section 4 presents the new broadcast algorithm for the mesh. Section 5 compares the performance of the proposed algorithm to the Recursive Doubling, Extending Dominating Node algorithms and Network Partitioning algorithms. Finally, Section 6 concludes this study.

2. CODED PATH ROUTING (CPR)

This section describes briefly the Coded Path Routing (CPR) approach that can reduce the overhead due to the start-up latency and the effects of the network size on the performance of collective communication. The CPR exploits the main features of wormhole switching, such as few buffer requirements and distance insensitivity, to overcome the limitations of the existing approaches, and to efficiently support collective communications. In the CPR, the header flit has two bits that form the control field. The two bits indicate to a router which action to take, e.g., pass or receive, upon the reception of a message.

```
Procedure Control Field (message, operation)

Begin

receive the second field of the message;

if (current router is not the addressed router) then

if (control field =10) then pass the message to the next router;

else

if (control field =01) then receive the message;

else

if (control field =11) then

{

receive the message; pass the message to the next router ;

}

else receive the message;

End
```

Figure 1. The "Control Field" algorithm used in the CPR.

Figure 1 describes the "Control Field" algorithm that the router use to either interpret or modify the control field. In fact, the two bits of the control field have originally been specified in order to enable the CPR to be used in different systems, such as those using one-port or multiple port router models, and also to support different types of collective communication operations, including broadcast and multicast. However, to illustrate the advantages of the CPR, we will focus our discussion in the present study on the use of the CPR for the development of broadcast algorithms; we plan to extend in the future the application of the CPR to multicast communication. As in the path-based algorithms of [12], which use the multidestination approach, it is assumed that a router in the CPR can simultaneously receive a message and passes a copy to the next router. Due to space limitation, we refer the reader to [1] for more detail on the CPR.

3. THE SYSTEM MODEL

A 2-dimensional mesh has $N = N_x \times N_y$ nodes, arranged in the two dimensions X, and Y respectively, with N_x and N_y being the number of nodes in the two dimensions. A node is identified by a two co-ordinate vector (x, y,), $0 \le x \le N_x - 1$, $0 \le y \le N_y - 1$. The mesh topology is asymmetric due to the absence of the wraparound connections along each dimension. Therefore, nodes may not be connected to the same number of neighbours; those at the corners, edges, and middle of the network have 2, 3 and 4 neighbours respectively. There are four corners in the mesh are located at the following addresses: $(N_x - 1, N_y - 1)$, $(N_x - 1, 0)$, $(0, N_y - 1)$, (0, 0). The discussion can be easily extended to

nodes situated at the corners and edges. A node consists of a processing element (PE) and router. The PE contains a processor and some local memory. A node uses four input and four output channels to connect to its neighbouring nodes; two in a dimension, one for each direction. There are also local channels used by the PE to inject/eject messages to/from the network, respectively. Messages generated by the PE are injected into the network through the injection channel. Messages at the destination node are transferred to the PE through the ejection channel. Similar to the previous studies of [8, 9], this study considers the multiple-port router model where multiple copies of the broadcast message can be injected into the network through different output channels concurrently. Furthermore, multiple broadcast messages can be transferred to the local PE.

4. THE NEW BROADCAST ALGORITHM

This section introduces the Parallel Coded Path (PCP for short) broadcast algorithm for multiple-port 2D mesh based on the CPR approach. A full description of the PCP algorithm is provided in Figure 2. For clarity, Figure 3 describes the operations of the proposed algorithm. The PCP algorithm exploits the features of the CPR to implement broadcast operations in two message-passing steps, thus considerably reducing the effects of the start-up latency. Examining Figures 2 and 3 shows that the PCP algorithm can achieve a high degree of parallelism during the propagation of the broadcast message from one router to the next. This has the net effect of greatly reducing the overall time required to complete a broadcast operation. Figure 3 illustrates two possible cases in which two source locations in a 10×8 mesh network. For each case, the PCP can be implemented with only two message-passing steps.

<u>Algorithm</u>: Broadcast in multiple-port 2-D Mesh. /* Input: 2D mesh $(N_x \times N_y)$, source node (S_x, S_y) , node (x, y) and (M: Message */ /* Output: All nodes receive a copy of M */ Initialise the sets $S_{upper} = \phi, S_{lower} = \phi, S_{right} = \phi, S_{left} = \phi$; Let $S_{corners}$ includes the four corner nodes in the mesh

 $S_{upper} = \{(x, y), 0 \le x < N_x, N_y / 2 \le y < N_y\},\$ $S_{lower} = \{(x, y), 0 \le x < N_x, 0 \le y < N_y / 2\}$
$$\begin{split} S_{right} &= \{(x, y), N_x / 2 \le x < N_x, 0 \le y < N_y\} \\ , S_{left} &= \{(x, y), 0 \le x < N_x / 2, 0 \le y < N_y\} \\ control field=00; \end{split}$$

 $if(S_x, S_y) \in S_{corners}$

Ł select two sides A and B := select(side1, side2, side3, side4): $\forall (x, y) \in A \text{ and } \forall (x, y) \in B \text{ in parallel do}$ 1 if $(x, y) \in A$ { Control field:=11; receive M; pass M to A; else control field :=10; while $(x, y) \notin S_{corners}$ {send M to the nearest corner;}; control field =11; receive M; pass M to B; select two sets S_1 , S_2 := select (Supper, Slower, Sright, Sleft); $\forall (x, y) \in A \text{ and } \forall (x, y) \in B \text{ in parallel do } \{$ control field :=11; $\forall (x, y) \in S_1 \text{ and } \forall (x, y) \in S_2$ receive M; pass M in S_1, S_2 else{ control field:=10; select two corners C_1, C_2 ; select two sides A and B; $\forall C_1 and C_2 in parallel do$ send M to C_1 and C_2 ? select two sets S_1 , S_2 ; $\forall (x, y) \in A \text{ and } \forall (x, y) \in B \text{ in parallel do } \{$ control field :=11; $\forall (x, y) \in S_1 \text{ and } \forall (x, y) \in S_2$ receive M; pass M in S_1 , S_2 } J end.

Figure 2. A description of the proposed PCP broadcast algorithm in the multiple-port 2-D mesh.

Theorem 1: The PCP broadcast algorithm in the 2D multiple-port mesh can be implemented in only two message-passing steps from any arbitrary source node, irrespective of the network size.

Proof: Consider the 2D mesh, with $N = N_x \times N_y$ nodes, and choose an arbitrary node as a source node (S_x, S_y) . According to the rules of the algorithm, depicted in Figs. 3.a and 3.b, respectively, the network is divided into four main sets: $(S_{upper}, S_{lower}, S_{left}, S_{left})$, where

$$\begin{split} S_{upper} &= \{(x,y), 0 \leq x < N_x, N_y / 2 \leq y < N_y \} \\ S_{lower} &= \{(x,y), 0 \leq x < N_x, 0 \leq y < N_y / 2 \} \;, \end{split}$$

$$\begin{split} S_{left} &= \{(x, y), 0 \leq x < N_x \, / \, 2, 0 \leq y < N_y \} \\ S_{right} &= \{(x, y), N_x \, / \, 2 \leq x < N_x, 0 \leq y < N_y \} \; . \end{split}$$

Based on the location of the source node, two sets of the nodes, row partitioning set and column partitioning set, can be considered simultaneously during broadcasting. Without loss of generality, let the source location (S_x, S_y) be as shown in Figure

3.c and Figure 3.d. The message is sent first to the nearest side of the mesh using any base routing, e.g. the dimension ordered routing, during the first message-passing step. Simultaneously, the opposite side sends a copy of the message exploiting the multiple port facility of the system. Each intermediate node receives and concurrently passes the message to the next node along the path as shown in Figures 3.c and 3.f. For each chosen side, when all the nodes have completely received the message, they concurrently send it to the rest nodes of their partition making the second message-passing step. \Box

Theorem 2: The PCP algorithm is deadlock free in the mesh.

Proof: We need to show that no cycle can occur during a message-passing step [2]. Let (S_x, S_y) be the source node initiating the broadcast operation. The target system can be divided into four main areas.

 $A_{x,y}$, where The first атеа $0 \le x < N_{\tau}$ and $y \in \{0, N_{y}\}$ includes the nodes of the two vertical sides of the system whereas the area $B_{x,y}$, where $x \in \{0, P_x\}$ and $0 \le y < N_y$ represents the two horizontal sides. Based on the rules of our algorithm and examining Figures 3.c and 3.d, which provide a description of the PCP algorithm, we show that the message can reach each area via either X or Y dimension, respectively. When the message reaches the third and fourth areas, $C_{x,y}$ and $D_{x,y}$, respectively, which of each equals to $C_{x,y}, D_{x,y} = \begin{cases} N/2 - A_{x,y} & \text{if the } A_{x,y} \text{ has been selected} \\ N/2 - B_{x,y} & \text{if the } B_{x,y} \text{ has been selected} \end{cases}$

In the second message-passing step, similar arguments can be used to show that there is no overlapping of the message paths can occur while broadcasting as illustrated in Figures 3.e and 3.f, respectively. \Box

5. PERFORMANCE RESULTS

We define the communication latency for a broadcast operation as the time when a broadcast message is injected into the network until the last node in the network receives the message. Let us now compare the performance of the proposed PCP algorithm to the well-known Recursive Doubling [2], Extended Dominating Node [9] and Network Partitioning algorithms [8] in terms of scalability, i.e. the number of message passing steps required to implement broadcast operation in different network sizes. We will use the short abbreviation PCP, RD, EDN and NP-D to refer to the four algorithms, respectively. Firstly, we will compare these broadcast algorithms in terms of the number of message-passing steps required. We then present results from simulation experiments to study the dynamic behaviour of the PCP under different traffic conditions. The RD was originally proposed by Barnett *et al* [2].

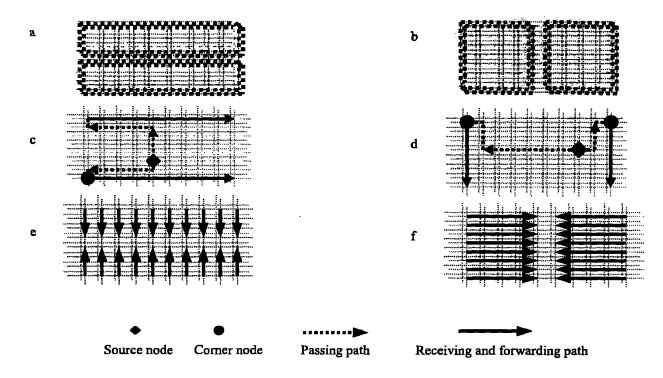


Figure 3. The two message-passing steps of the PCP in different possible cases. (a) a horizontal partition case, (b) a vertical partition case, (c), (e) the two steps in the horizontal partitioning case and (d), (f) the two steps in the vertical partitioning case.

This algorithm requires Log_2N steps for broadcasting in 2D mesh. In this algorithm, each node holding a copy of the message is responsible for a partition of a row or column, which will be then divided in half. In each half, a node sends a copy of the message to the node in the other half that occupies the same relative position. The EDN was proposed by Tsai and McKinley [9]. The authors in [9] have shown that the number of message-passing steps required in a network size of $(2^k \times 2^k)$ is k+1. The EDN scheme is based on the concept of dominating sets of Graph Theory. In [8], the NP-D has been proposed to achieve more parallelism during broadcast operation. The broadcast algorithm works in three stages. First, the main network is divided into sub-networks and the source node divides the message M

broadcast into sub-messages h; a leader node is chosen for each sub-network. Second, the leader nodes perform a broadcast operation concurrently. Third, the sub-networks collect sub-messages $M_0, M_1, ..., M_{h-1}$ from the nodes that have received sub-messages in the second stage and combine them into the main message M [10].

However, the NP-D does not minimise the message-passing steps required to perform broadcast operation in that it requires k+2 in order to perform a broadcast operation in the $(2^k \times 2^k)$ mesh [10]. Figure 4 plots the number of message-passing steps required for a broadcast operation by the four algorithms in the $(2^k \times 2^k)$ mesh. For the sake of illustration, the network size was varied from 64 to 20376 nodes. The results reveal that while the number of steps required by the EDN, RD and NP-D algorithms increases with the system size, that required by the PCP is fixed (only two) regardless of the network size.

Having examined the improvement in broadcast achieved by the PCP in the terms of the number of message-passing steps required, in what follows we conduct a performance comparison of the PCP and RD under dynamic situations using simulation experiments. We have decided to exclude the EDN and NP-DA algorithms from the comparison because the RD is much more able to take advantage of the pipelining effect of wormhole routing to avoid channel contention among messages [9]. It is worth mentioning that the values of the communication latency parameters are consistent with the Cray T3D (channel rate was set at $\beta = 0.0033 \,\mu s$ and $\alpha = 0.75 \,\mu s$) [9].

Figure 5 plots the communication latency versus different network sizes with 100 flits for the message length. The results from the figure show that the advantage of our algorithm is significant as it requires a fixed number of message-passing steps in the all network sizes and it is able to fully utilise the multipleport facility of the system. In contrast, Figure 5 confirms the fact that the RD performance is highly dependent on the network size. In Figure 6, we have examined the performance of the CPR and the RD using two network sizes $N=4 \times 4$ and 8×8 . The message length was varied from 30 to 210 flits and the parameters of the communication latency were set in a way to be consistent with the Cray T3D [1].

The figure reveals that there is a slight difference in the achieved performance of the PCP in the two network sizes. This is because in the PCP, when the network size increases, only the distance (i.e., the number of nodes) traversed by the message increases while there is no increase in the number of message-passing steps required to complete the broadcast operation. Taking the advantage of wormhole distance insensitivity, the PCP algorithm has also become insensitive to the network size. However, turning to the RD shows the high effect of network size on the algorithm performance. If we increase the network size to 8×8 , the advantage of the CPR becomes more noticeable. This is due to the fact that the RD, like most of existing algorithms, is highly sensitive to the number of the message-passing steps.

required to meet the increase in the network size, and the increase in the message length.

6. CONCLUSION

There has been hardly any attempt to provide broadcast algorithms with high scalability and parallelism. In an effort to fill this gap, this paper has presented a new efficient algorithm. The proposed algorithm has the main advantage of exhibiting a high degree of parallelism and requiring a fixed number of messagepassing steps irrespective of the network size. Furthermore, a performance analysis has revealed that the proposed algorithm has

7. REFERENCES

- A. Y. Al-Dubai, M. Ould-Khaoua, Coded path routing: A new approach to broadcasting in 3-D meshes, Proc. 20th IEEE Int. Performance, Computing and Communications Conference (IPCCC'2001), IEEE Computer Society Press, Arizona, USA, pp 155-162, April 4-6, 2001.
- [2] M. Barnett, G. David, R. A. van de Geijn, J. Watts, Broadcasting on meshes with wormhole routing, J. Parallel & Distributed Computing, vol. 35, 111-122, 1996.
- [3] J. Duato, S. Yalamanchili, L. Ni, Interconnection networks: An engineering approach, *IEEE Computer* Society Press, 1997.
- [4] P. K. McKinley, H. Xu, A. Esfahanian, L. M. Ni, Unicastbased multicast communication in wormhole-routed direct networks, *IEEE TPDS*, vol. 5, no. 12, 1254-1265, 1994.
- [5] T.-S. Chin, C.-Y. Chang, J-P. Sheu, Efficient path-based multicast in wormhole-routed mesh networks, J. Systems Architecture, vol. 46, 919-930, 2000.
- [6] L. Ni, P. McKinley, A survey of wormhole routing techniques in direct networks, *Computer*, vol. 26, no. 2, pp. 62-76, Feb. 1993.
- [7] D. K. Panda, S. Singal, R. Kesavan, Multidestination message-passing in wormhole k-ary n-cube networks with base routing conformed paths, *IEEE TPDS*, vol. 10, no. 1, pp. 76-96, 1999.
- [8] Y.-C. Tseng, S.-Y. Wang, C.-W. Ho, Efficient broadcasting in wormhole-routed multicomputers: A network-partitioning approach, *IEEE TPDS*, vol. 10, no. 1, pp. 44-61, 1999.
- [9] Y.-J Tsai, P.K. McKinley, An extended dominating node approach to broadcast and global combine in multiport wormhole routed mesh networks, *IEEE TPDS*, vol. 8, no. 1, 41-58, 1997.
- [10] S. Cang, J. Wu, Time-step optimal broadcasting in 3-D meshes with minimal total communication distance, J. Parallel & Distributed Computing, vol. 60, pp. 966-997, 2000.
- [11] J. Watts, Efficient collective communication on multidimensional meshes with wormhole routing. Tech. Rep. TR-94-19. Dept. Computer Science, Univ. Texas at Austin, June 1994.
- [12] D. F. Robinson, P. K. McKinley, C. Cheng, Path based multicast communication in wormhole routed unidirectional torus networks, JPDC, vol. 45, 104 - 121, 1997.

superior latency characteristics the existing Recursive Doubling, Extending Dominating Nodes and Network Partitioning algorithms. The next step in our work is to devise new multicast algorithms using the CPR and compare their performance with existing algorithms. Another possible line for future research is to extend the CPR to support collective communication in other common multicomputer networks, such as the torus and hypercube.

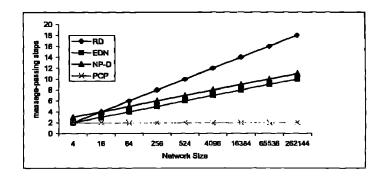


Figure 4. Comparison of the number of message passing steps (Scalability)

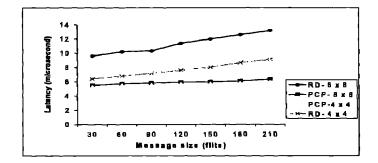


Figure 5. Comparison of the broadcast algorithms in different network sizes (with the Cray T3D parameters).

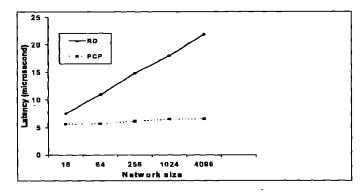


Figure 6. Comparison of the broadcast algorithms in the 8×8 mesh (with the Cray T3D parameters).