On Single-Port Multinode Broadcasting

Vassilios V. Dimakopoulos
Department of Computer Science, University of Ioannina
P.O. Box 1186, GR-45110 Ioannina, Greece
E-mail: `dimako@cs.uoi.gr`

Abstract

Multinode broadcasting, an important collective communication problem, involves simultaneous broadcastings from all the nodes in a network. In this work we present algorithms for the minimum-time solution of the problem in packet-switched networks that follow the single-port model. In particular, we construct a general algorithm for the solution of the problem in arbitrary multidimensional networks and provide conditions that ensure its optimality.

1 Introduction

The advent of distributed-memory multiprocessors has spawned an increasing amount of research in information dissemination problems. Given a network of processors (or *nodes*) where some of them own pieces of information, the problem is to spread the information to a group of recipients using the links of the network. The term 'collective communications' has been coined to signify the fact that such problems involve more than two nodes.

Among the variety of situations that can arise in practice (e.g. in parallel numerical algorithms [2]) broadcasting and multinode broadcasting intrigued the research community as early as the 1950's [7]. In *broadcasting* there is one node that owns a piece of information (hereafter called 'message' or 'packet') and needs to send it to the other nodes in the network. *Multinode broadcasting*, which is the subject of this paper, involves simultaneous broadcastings from all nodes, that is, every node has to send a single message to all the other nodes in the network.

Two surveys on collective communication problems, including multinode broadcasting, were given in [7, 6]. Multinode broadcasting is also known as *all-to-all broadcasting* and *gossiping*. Traditionally, though, the term 'gossiping' implies certain assumptions about the communication cost. In particular, it is assumed that whenever two neighboring nodes communicate (known as a 'call') they can exchange any number of packets in one time unit. Such a model is not suited for our purposes here; we are interested in obtaining the minimum possible *time* needed to solve the problem, instead of the minimum number of calls. Clearly, in practice the more packets two nodes exchange the more the time delay will be. As a consequence, we will make the more realistic assumption that in one time unit a node can only send one packet.

We consider packet-switched networks that follow the constant communication paradigm [6] where: (a) communication links are bidirectional, (b) a message requires one time unit (or *step*) to be transfered between two nodes and (c) only adjacent nodes can exchange messages. Furthermore, nodes will be assumed to have *single-port* capabilities, that is, a node can only communicate with only one neighbor at a time. Under such a scheme two basic possibilities arise:

- the *SAR model* where a node can *s*end a message *and* simultaneously *r*eceive a message in a step
- the *SOR model* where a node can *s*end a message to *or* *r*eceive a message from a neighbor but not both simultaneously.

Notice that in the SAR model a node may send a message to some neighbor but receive a message possibly from a different neighbor. This is more general than the so-called 'telephone' model where message transmission and reception occurs to/from the same neighbor. The latter was assumed recently in the work of J.-C. Bermond et al [1]. The SOR model is also known as the 'telegraph' model.

We study the multinode broadcasting problem for both models since they yield different solutions. For each of the models we derive simple lower bounds on the time needed to complete a multinode broadcasting operation. We then proceed to construct a general algorithm that solves the multinode broadcasting problem in *any* multidimensional (or cartesian product) network in a modular way. Such networks are probably the most popular as is evident from their utilization in commercial parallel machines (e.g. Cray T3D, Intel Paragon). Assuming we are given algorithms for each of the dimensions, we show how to construct an algorithm for the multidimensional graph. We also prove that if the algorithms for each of the dimensions are time-optimal, then the derived algorithm is also optimal. An analogous theory was previously developed for the total exchange problem [5].

2 Lower Bounds

Let us consider a network (or graph) $G = (V, E)$ where V is the node (or vertex) set and E is the link (or edge) set, and let $n = |V|$. J.-C. Bermond et al [1] derived quite tight bounds for the time needed to perform multinode broadcasting in arbitrary networks, using the telephone model. We derive here similar, and simpler bounds for the SAR and SOR models.

In the multinode broadcasting problem each node will receive $n-1$ messages, one from each of the other nodes. Under the single-port assumption messages at any node can only arrive one by one. If T_{SAR}, T_{SOR} are the times needed to perform multinode broadcasting in G under the corresponding models, we have the lower bound of:

$$T_{SAR}, T_{SOR} \geq n-1 \quad \text{steps.} \tag{1}$$

For the case of SOR model the lower bound can be further tightened. Each of the n broadcast messages must be received by $n-1$ nodes. In other words, each message implies $n-1$ receptions. For the receptions to occur, there must clearly occur $n-1$ transmissions, too. In total, for all n messages there will occur:

$$n(n-1) \text{ receptions and } n(n-1) \text{ transmissions.}$$

If the number of nodes is *even* then at most n actions (receptions or transmissions) can occur at each step. To make it clearer, since no node is allowed to simultaneously receive and transmit messages, at most half of the nodes can send a message and at most half of the nodes can receive a message. As a result, at most $n/2$ transmissions and $n/2$ receptions may be had at each step. This gives the lower bound of:

$$T_{SOR}^{even} \geq 2(n-1). \tag{2}$$

If n is odd then at most $n-1$ actions (receptions or transmissions) can occur at each step, that is, one node can not participate at all, giving the lower bound of:

$$T_{SOR}^{odd} \geq 2n. \tag{3}$$

3 Multinode Broadcasting in Multidimensional Networks

In this section we are going to develop a general multinode broadcasting algorithm for *any* multidimensional graph. Although simple, the algorithm will be shown to be optimal if certain conditions are met.

Given k graphs $G_i = (V_i, E_i)$, $i = 1, 2, \ldots, k$, their (cartesian) product is defined as the graph $G = G_1 \times \cdots \times G_k = (V, E)$ whose vertices are labeled by a k-tuple $(v_1, \ldots, v_k)$ and

$$V = \Big\{ (v_1, \ldots, v_k) \mid v_i \in V_i, i = 1, \ldots, k \Big\}$$
$$E = \Big\{ \big((v_1, \ldots, v_k), (u_1, \ldots, u_k)\big) \mid \exists j \text{ s.t. } (v_j, u_j) \in E_j \text{ and } v_i = u_i \text{ for all } i \neq j \Big\}.$$

We will call such products of graphs *multidimensional* graphs and G_i will be called the ith *dimension* of the product. The ith component of the address tuple of a node will be called the ith *address digit* or the ith *coordinate*. The definition of E above in simple words states that two nodes are adjacent if they differ in exactly one address digit. Their differing coordinates should be adjacent in the corresponding dimension. An example is given in Fig. 1.

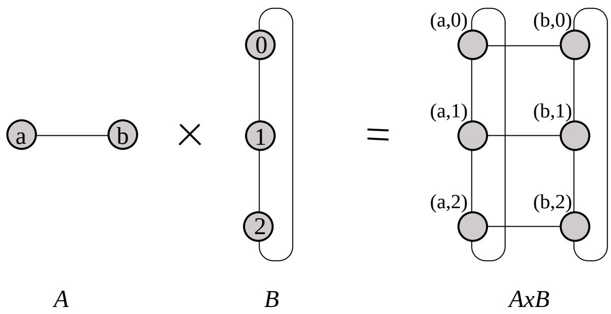

Figure 1: A two-dimensional graph

Hypercubes are products of two-node linear arrays (or rings), tori are products of rings. If all dimensions of the torus consist of the same ring, we obtain k-ary n-cubes [4]. Meshes are products of linear arrays [8]. Generalized hypercubes are products of complete graphs [3]. Multidimensional graphs have $n = n_1 n_2 \cdots n_k$ nodes, where $n_i = |V_i|$ is the number of nodes in G_i, $i = 1, 2, \ldots, k$.

It will be convenient to use the *don't care* symbol '$*$' as a shorthand notation for a set of addresses. An appearance of this symbol at an element of an address tuple represents all legal values of this element. In the last example, $(a, *) = \{(a, 1), (a, 2), (a, 3)\}$, $(*, 1) = \{(a, 1), (b, 1)\}$ while $(*, *)$ denotes the whole node set of the graph.

3.1 The Algorithm

Let $G = A \times B$. A k-dimensional network $G_1 \times \cdots \times G_k$ can still be expressed as the product of two graphs by taking $A = G_1$ and $B = G_2 \times \cdots \times G_k$, so we may consider two dimensions without loss of generality. Let

$A = (V_A, E_A)$, $B = (V_B, E_B)$, $G = (V, E)$, $n_1 = |V_A|$, $n_2 = |V_B|$ and $n = n_1 n_2$. Finally, let:

$$\begin{aligned} V_A &= \{v_i \mid i = 1, 2, \ldots, n_1\} \\ V_B &= \{u_i \mid i = 1, 2, \ldots, n_2\}. \end{aligned}$$

Graph G can be viewed as n_2 (interconnected) copies of A. For example, in Fig. 1 graph $A \times B$ consists of three copies of A, where the corresponding nodes are interconnected according to the edges in B. Let A_j be the jth copy of A with node set $(*, u_j)$, where $*$ takes all values in V_A. Similarly, G can be viewed as n_1 copies of B, and we let B_i be the ith copy of B with node set $(v_i, *)$.

Let us assume that there exist multinode broadcasting algorithms for each of the dimensions. Our method utilizes the algorithms for the dimensions so as to synthesize an algorithm for the whole graph. In other words, the problem of performing multinode broadcasting in $G = A \times B$ is decomposed to the simpler problem of performing multinode broadcasting in A and in B. This is a highly desirable simplification since multidimensional networks are quite complex structures.

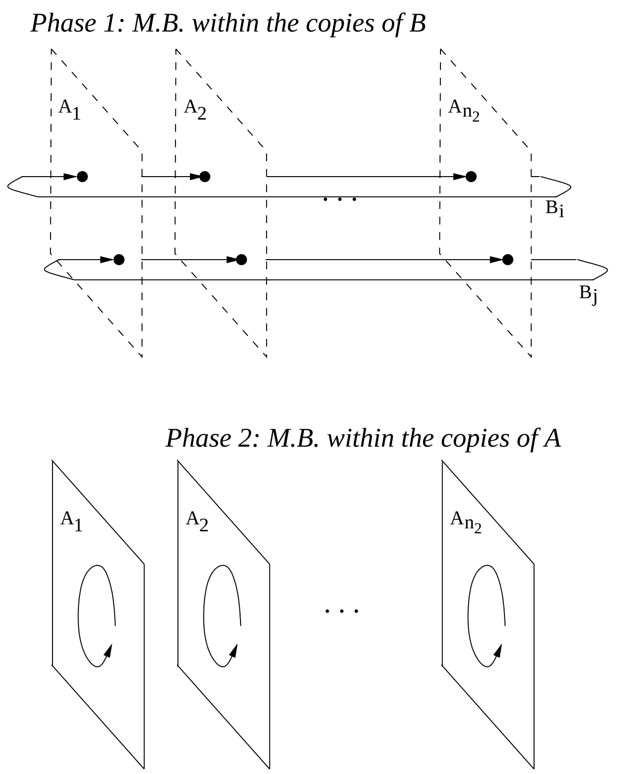

Figure 2: The two phases of the multinode broadcasting algorithm

The algorithm we are going to give for G consists of two phases, and is illustrated in Fig. 2. In Phase 1 nodes perform a multinode broadcasting in the second dimension. Notice that because there are no nodes in common between the copies of B what we shall describe for a certain copy B_i occurs simultaneously for the other of copies of B, too.

Consider node (v_i, u_j) in B_i, and let $m(v_i, u_j)$ denote its own broadcast message. When Phase 1 is complete, this node will contain all broadcast messages from the nodes in B_i, that is, it will contain the messages of nodes $(v_i, *)$: $m(v_i, u_1), m(v_i, u_2), \ldots, m(v_i, u_{n_2})$. This holds for every node in the graph. In particular, every node in A_j will contain n_2 messages, each one originating from a different node. In other words, at the end of Phase 1 all $n = n_1 \times n_2$ broadcast messages will have been received by the n_1 nodes in A_j (for all j), each node holding n_2 of them.

It should now follow clearly that the way to distribute those messages to every node in A_j is through multinode broadcastings within A_j. Thus Phase 2 consists of a series of n_2 multinode broadcastings in the first dimension in order for each node to broadcast the n_2 messages it holds.

Phase 1:
1 Do in parallel for all B_i, $i = 1, 2, \ldots, n_1$
2   In B_i perform multinode broadcasting;

Phase 2:
3 For $r = 1, 2, \ldots, n_2$
4   Do in parallel for all A_j, $j = 1, 2, \ldots, n_2$
5   In A_j perform multinode broadcasting with
          node (v_i, u_j) broadcasting $m(v_i, u_r)$ $(i = 1, 2, \ldots, n_1)$;

Figure 3: The algorithm for graph $G = A \times B$

The algorithm is summarized in Fig. 3, and is a general solution to the multinode broadcasting problem for any multidimensional network. If the network has $k > 2$ dimensions $G = G_1 \times \cdots \times G_k$, the algorithm can be used recursively by taking $A = G_1$ and $B = G_2 \times \cdots \times G_k$. Phase 1 (lines 1–2) can be performed by invoking the algorithm with $A = G_2$ and $B = G_3 \times \cdots \times G_k$ and so forth.

Notice that the algorithm is independent of the link model in use. Because the two phases are executed one after the other, and within Phase 2 the multinode broadcastings are also executed serially, only one multinode broadcasting operation is in effect at any step. As a result, the whole algorithm is consistent with the link model of the algorithms for each dimensions. For example, if the algorithms for A and B operate on SOR networks, so does our algorithm for $G = A \times B$.

3.2 Optimality conditions

We proceed now to determine the time required for the general algorithm in Fig. 3 and the conditions under which it behaves optimally. Let T_A and T_B denote the number of steps needed to perform multinode broadcasting in A and B correspondingly.

Theorem 1 *The multinode broadcasting algorithm for $G = A \times B$ requires:*

$$T = T_B + n_2 T_A \quad \textit{time units.}$$

Proof. The result is straightforward: Phase 1 performs multinode broadcasting within B_i (for all $i = 1, 2, \ldots, n_1$ in parallel), taking thus time equal to T_B. Phase 2 performs n_2 multinode broadcasting operations within A_j (for all $j = 1, 2, \ldots, n_2$ in parallel), each requiring T_A steps. $\square$

Theorem 2 *Under the SAR model, if multinode broadcasting in A and B can be performed in time equal to the lower bound of Eq. (1) then the same is true for $G = A \times B$.*

Proof. If T_A and T_B achieve the lower bound of Eq. (1) then $T_A = n_1 - 1$ and $T_B = n_2 - 1$. From Theorem 1 we obtain:

$$T = (n_2 - 1) + n_2(n_1 - 1) = n_1 n_2 - 1 = n - 1,$$

as required. $\square$

Theorem 3 *Under the SOR model, if both dimensions have an even number of nodes and multinode broadcasting in each one can be performed in time equal to the lower bound of Eq. (2) then the same is true for $G = A \times B$.*

Proof. If both dimensions have an even number of nodes then by Eq. (2) we must have $T_A = 2(n_1 - 1)$ and $T_B = 2(n_2 - 1)$. From Theorem 1 we obtain:

$$T = 2(n_2 - 1) + 2n_2(n_1 - 1) = 2n_1 n_2 - 2 = 2(n - 1),$$

which is optimal. $\square$

Theorem 4 *Under the SOR model, if only one dimension has an even number of nodes and multinode broadcasting in each dimension can be performed in time equal to the lower bound of Eq. (2) or Eq. (3) then the algorithm for $G = A \times B$ is optimal within two steps.*

Proof. Without loss of generality assume that n_2 is odd and n_1 is even. Otherwise, we might rename the dimensions by considering graph $G = B \times A$ which is isomorphic to $A \times B$. Thus, according to Eqs. (2) and (3) we must have $T_A = 2(n_1 - 1)$ and $T_B = 2n_2$. From Theorem 1 we obtain:

$$T = 2n_2 + 2n_2(n_1 - 1) = 2n_1 n_2 = 2n.$$

Since n is even, the lower bound of Eq. (2) shows that T is suboptimal by 2 steps. $\square$

4 Conclusion

We have considered the problem of multinode broadcasting in packet-switched networks with single-port capabilities, whereby a node can at most send and/or receive one message at each step. We note that this is exactly what most present-day machines are capable of. Under this model, we consider the case where simultaneous transmission and reception is allowed at each node (SAR) and the case where a node is only allowed to either transmit or receive at each step (SOR).

We provided a general solution to the problem for multidimensional networks. We derived a modular method that utilizes multinode broadcasting algorithms for each of the dimensions. Our scheme is simple but it nevertheless maintains optimality always under the SAR model and in most cases under the SOR model. An interesting course of further research is the case of the SOR model where all dimensions of the network contain an odd number of nodes. In this setting our method cannot maintain optimality even if multinode broadcasting in every dimension can be performed optimally.

References

[1] J. - C. Bermond, L. Gargano, A. A. Rescigno and U. Vaccaro, "Fast gossiping by short messages," *SIAM Journal on Computing*, Vol. 27, No. 4, pp. 917–941, Aug. 1998.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewoods Cliffs, N.J.: Prentice - Hall, 1989.

[3] L. N. Bhuyan and D. P. Agrawal, "Generalized hypercube and hyperbus structures for a computer network," *IEEE Trans. Comput.*, Vol. C-33, No. 4, pp. 323–333, Apr. 1984.

[4] W. J. Dally and C. L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," *IEEE Trans. Comput.*, Vol. C-36, No. 5, pp. 547–553, May 1987.

[5] V. V. Dimakopoulos and N. J. Dimopoulos, "A theory for total exchange in multidimensional interconnection networks," *IEEE Trans. Parall. Distrib. Syst.*, Vol. 9, No. 7, pp. 639–649, July 1998.

[6] P. Fraigniaud and E. Lazard, "Methods and problems of communication in usual networks," *Discrete Appl. Math.*, Vol. 53, pp. 79–133, 1994.

[7] S. M. Hedetniemi, S. T. Hedetniemi and A. L. Liestman, "A survey of gossiping and broadcasting in communication networks," *Networks*, Vol. 18, pp. 319–349, 1988.

[8] F. T. Leighton, *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*. San Diego, CA: Morgan Kaufmann, 1992.