Broadcasting in WDM Optical Rings and Tori

Bruno Beauquier

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Broadcasting in WDM Optical Rings and Tori

Bruno Beauquier

N˚ 3410

Avril 1998

THÈME 1

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

Broadcasting in WDM Optical Rings and Tori

Bruno Beauquier*

Thème 1 — Réseaux et systèmes
Projet Sloop

Rapport de recherche n° 3410 — Avril 1998 — 15 pages

Abstract: The well-known spanning binomial tree broadcast algorithm is generalized to obtain two families of broadcast algorithms for optical rings and two-dimensional toroidal meshes (tori) using *Wavelength Division Multiplexing (WDM)*. These generalizations take advantage of the concurrent transmission through optical links offered by WDM. Their performances are measured under the *linear cost model* : the cost of sending a message of L bits is defined as $\alpha + L\tau$, where α is the latency and τ is the per-byte transmission cost. It is assumed that each node can concurrently transmit one message and receive one message. Our algorithms are based on the familiar spanning binomial tree and on the dimensional exchanges commonly used on hypercubes. We restrict the number of nodes in a ring and in each dimension of a torus to be a power of two. The algorithms described in this paper offer significant performance improvements over the basic spanning tree broadcast.

Key-words: Optical networks, WDM, broadcasting.

To appear in the DIMACS Series on Discrete Mathematics and Theoretical Computer Science published by the American Mathematical Society. Presented in the workshop on "Multichannel Optical Networks: Theory and Practice", (DIMACS, March 16-19, 1998).

* SLOOP, joint project I3S-CNRS/UNSA/INRIA. Work partially supported by the AFIRST in the framework of the French-Israeli project *Communication Algorithms in Optical Networks*. Email : beauquier@sophia.inria.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : 04 92 38 77 77 - International : +33 4 92 38 77 77 — Fax : 04 92 38 77 65 - International : +33 4 92 38 77 65

Diffusion par multiplexage en longueur d'onde dans les anneaux et les tores optiques

Résumé : L'algorithme bien connu de diffusion sur un arbre couvrant binômial est généralisé pour obtenir deux familles d'algorithmes de diffusion dans des anneaux et des tores optiques utilisant la technique du *multiplexage en longueur d'onde* (en anglais, *Wavelength Division Multiplexing* : WDM). Ces généralisations exploitent les avantages de la technologie WDM qui permet des transmissions de données concurrentes au sein d'une même fibre optique. Les performances des algorithmes sont mesurées par un modèle de coût affine : le temps pour envoyer un message de L bits est défini par $\alpha + L\tau$, où α est la latence et τ l'inverse de la bande passante. Il est supposé que chaque nœud peut en même temps envoyer et recevoir un seul message. Nos algorithmes sont basés sur des plongements des arbres couvrants binômiaux et des échanges dimensionnels, classiquement utilisés dans les hypercubes.

Mots-clés : Réseaux optiques, multiplexage en longueur d'onde, diffusion.

1 Introduction

Optics is emerging as a key technology in communication networks, promising very high speed local or wide area networks of the future. A single optical wavelength supports rates of gigabits per second, which in turn support multiple channels of voice, data and video [4, 6]. Multiple laser beams that are propagated over the same fiber on distinct optical wavelengths can increase this capacity even further. This is achieved through *Wavelength Division Multiplexing* (or *WDM*) [3], by partitioning the optical bandwidth into several channels and allowing the transmission of multiple data streams concurrently along the same optical fiber.

The problem we consider here is motivated by *multi-hop* communication networks [7] with reconfigurable wavelength selective optical switches, without wavelength converters. Different signals may travel on the same fiber-optic link (but on different wavelengths) into a node, and then exit from it on different links, keeping their original wavelengths.

In such communication networks, each source-destination pair can be connected by an end-to-end transparent channel that is identified through a wavelength and a physical path. This necessitates to tune the sender and the receiver to the same wavelength and to set the intermediate optical switches in the right configuration. Wavelengths being a limited resource, solutions to the problem of efficient routing and wavelengths allocation are of importance for the future development of optical technology.

A *broadcast* is an operation where one node of a distributed multicomputer network has a message that must be copied to each other node. The work presented here offers solutions to the broadcast problem for optical rings and toroidal meshes based on the familiar spanning binomial tree and on dimensional exchanges commonly used on hypercubes [5].

The paper is organized as follows. The next section describes the communication model on which the analysis of communication algorithms will be based. The broadcast algorithms and their costs are given in Section 3. In Section 4 all the results are summarized and comments are made.

2 Communication Model

This work considers solutions to the broadcast problem on rings of $n = 2^d$ processors and on two-dimensional toroidal meshes of $n = n_1 \times n_2 = 2^{d_1} \times 2^{d_2}$ processors. The nodes of a ring are numbered by integers i modulo n. The nodes of a toroidal mesh (*torus*) are numbered by couples (i, j) of an integer i modulo n_1 and an integer j modulo n_2. Each node is connected by a pair of unidirectional fiber-optic links (one in each direction) to each of its immediate neighbours in the horizontal and vertical directions. Each node can concurrently transmit one message and receive one message. In this paper, a message in a torus is routed or only horizontally either only vertically to the receiver. Messages can be routed through a node without affecting its performance as a sender or receiver. It is assumed that all links have the same fixed bandwidth. *Wavelength conflict* occurs when the paths taken by two or more messages sent on the same wavelength have one or more links in common.

Finally, it is assumed that each receiving node allocates a buffer for each incoming message before it arrives. Under this assumption the sender can transmit a message without prior handshaking with the receiver.

A simple communication model describes the cost of sending a message of L bits as $\alpha + L\tau$, where α is the latency (including the *start-up*, the switching and the wavelength tuning delays) and τ is the per-byte transmission cost ($1/\tau$ corresponds to the bandwidth). A close examination of real message passing networks reveals a much larger collection of factors that can affect the cost of communication algorithms. Some of these factors are the length of the circuit over which the message travels, the effects of message packetization performed by the node operating system, the packet permutation costs within a node, the synchronization costs. The permutation costs generally affect the cost of global communication and remarks will be given in Section 4. The remaining factors, which have relatively small effects on cost, will not be considered here.

3 Broadcast Algorithms

Two families of broadcast algorithms for optical rings and toroidal meshes are considered. Each algorithm is based on communication patterns commonly used on hypercubes, such as spanning binomial trees and dimensional exchanges. It is assumed w.l.o.g. that node 0 of the ring (respectively node $(0,0)$ of the torus) contains a message of length L to be broadcast to all the other nodes. The *cost* of a broadcast operation is measured from the time the originator begins the broadcast to the time the last node receives the message.

The familiar spanning tree broadcast algorithm is considered first. Improvements of that basic algorithm will take advantage of 1) the concurrent transmission through optical links offered by WDM, 2) the additional bandwidth offered by bidirectional links, and 3) the increased connectivity of toroidal meshes over rings.

3.1 Basic subroutines

3.1.1 Spanning tree broadcast

The first algorithm is based on the familiar spanning binomial tree that is used in the recursive broadcast algorithm for hypercubes [5]. This algorithm will be called the *spanning tree (ST) broadcast* and was described earlier in [1] for linear arrays and meshes. A ST broadcast on a ring or a torus of 2^d nodes takes d rounds. On the i^{th} round ($1 \leq i \leq d$), each node j that already has a copy of the message sends it to node $j \oplus 2^{d-i}$, where $\oplus$ denotes bit-wise exclusive OR. This is illustrated for a ring in Figure 1(a). The corresponding spanning tree is shown in Figure 1(b). Each arc of the tree is labeled with the round at which it carries the message.

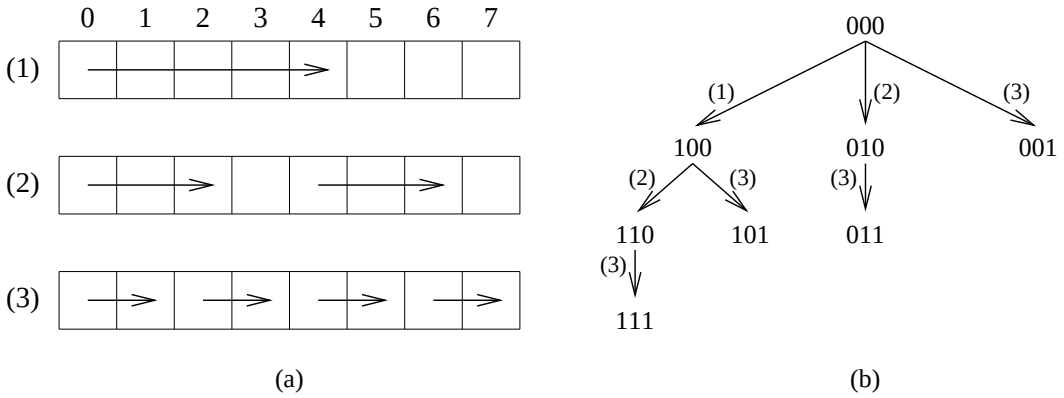

Figure 1: Spanning tree broadcast on a ring of 2^3 nodes.

The spanning binomial tree on which the ST broadcast algorithm is based is also used in all of the other broadcast algorithms considered here. The nodes of such trees are numbered in binary. (From a purely graph-theoretic point of view, these trees are not spanning trees, but they are useful for describing the scheduling and routing of messages and they will continue to be referred to as spanning trees here.) For purposes here, a *spanning tree with root 0* is a directed graph of $n = 2^d$ nodes in which each node j has children whose node numbers are obtained by complementing exactly one of the trailing zeros (if any) of j. To determine the node numbering of a spanning tree with root other than 0, exclusive OR the node number of each node in the tree with the node number of the root. For more details about the properties of these trees see [5].

In the spanning tree of a ring of 2^d nodes, some tree arcs represent directed paths of several links in the ring and some links are used in several tree arcs. Since some of the tree arcs carry messages simultaneously during this broadcast algorithm, there might be wavelength conflict. (This possibility does not arise in hypercube because there is a one-to-one correspondence between tree arcs and hypercube links.) Even though some arcs in the tree share the same links, at each round only disjoint sets of links are used, and so one wavelength is sufficient.

A ST broadcast on a torus of $2^{d_1} \times 2^{d_2} = 2^d$ nodes, based on the algorithm for rings, is shown in Figure 2. In this algorithm the message is first broadcast to the nodes in the row containing the originator, then each node in that row broadcasts the message to the nodes in its column. As in the case of the ring, wavelength conflict can be avoided with the use of one wavelength only.

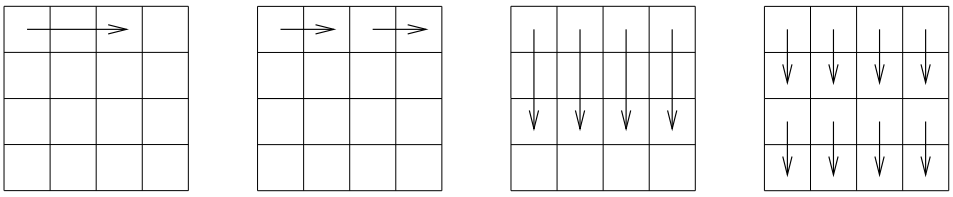

Figure 2: A simple spanning tree broadcast on a torus of $2^2 \times 2^2$ nodes.

Thus, for both a ring of 2^d nodes and a torus of $2^{d_1} \times 2^{d_2} = 2^d$ nodes, the ST broadcast algorithm uses one wavelength and has cost

$$d(\alpha + L\tau) \tag{1}$$

There are two other communication problems whose solutions are used frequently in the broadcast algorithms given here. Solutions to these two problems, based on spanning trees, are now described.

3.1.2 Distribute algorithm

Some of the algorithms that follow make use of the *distribute* operation in which one node sends a distinct message to each other node in the network. This operation is also called *scatter* or *one-to-all personalized communication* [5]. In all of the broadcast algorithms of Section 3.2, the messages to be distributed, called *packets*, will arise by partitioning the original message that is to be broadcast. In a ring or a torus of $n = 2^d$ nodes, each of the distributed packets, has length L/n, where L is the total length of the message to be broadcast. The distribute algorithm used here is based on a spanning tree : a message of length L is distributed to all the nodes in the tree by halving it at each round until each node has received its packet. For both a ring of 2^d nodes and a torus of $2^{d_1} \times 2^{d_2} = 2^d$ nodes, the cost of this distribute algorithm (using one wavelength) is

$$\sum_{i=1}^{d} \left(\alpha + \frac{L\tau}{2^i} \right) = d\alpha + \left(1 - \frac{1}{2^d} \right) L\tau \tag{2}$$

3.1.3 All-to-all broadcast

The other problem of interest is called the *all-to-all broadcast*. In this problem, *each* node has a message (typically, a packet of length L/n) that must be broadcast to all other nodes. This problem is easily solved by exchanging packets between nodes whose node numbers (written in binary) differ by one bit. On hypercubes this algorithm is known as a *dimensional exchange* and the same term will be used here. Note that the message length doubles at each round of this algorithm and so its cost is the same as the cost (2) of the distribute algorithm described above, provided that there is no wavelength conflict. However, on the

topologies considered in this paper a dimensional exchange can give rise to link contention, whose issue is addressed now.

In a ring of 2^d nodes, the all-to-all broadcast algorithm can be realized with 2^{d-2} wavelengths, that we number from 0 to $(2^{d-2}-1)$, by making node numbered j communicate with wavelength numbered $[j] = j \bmod 2^{d-2}$ (see Figure 3). On the first round, each node has to send its packet to the corresponding opposite node in the ring. No wavelength conflict occurs if half of the nodes using the same wavelength communicate in the clockwise direction of the ring, while the other half of nodes use the other direction.

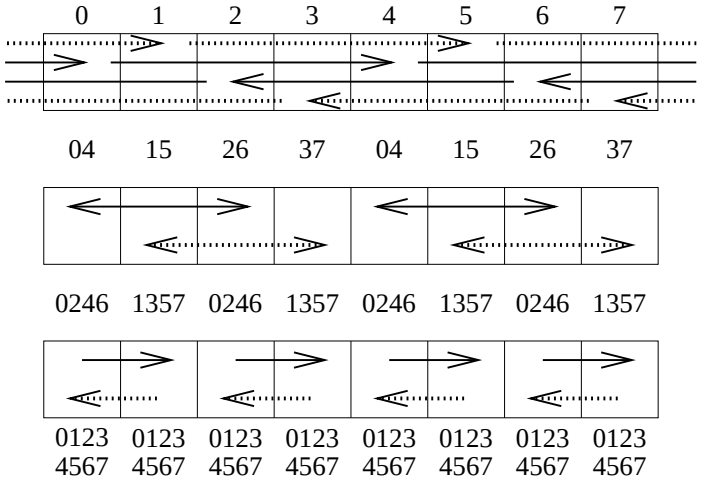

Figure 3: All-to-All broadcast algorithm using two wavelengths in a ring of 2^3 nodes.

In a torus of $2^{d_1} \times 2^{d_2}$ nodes (assuming that $d_1 \geq d_2$), the first $(d_1 - d_2)$ rounds of the all-to-all broadcast algorithm consist in exchanges of packets along the longest axis. As in the case of the ring, this can be done with 2^{d_1-2} wavelengths. Then it remains to realize a dimensional exchange in $2^{d_1-d_2}$ contiguous square blocks of $2^{d_2} \times 2^{d_2}$ nodes. Making half of the nodes communicate horizontally while the other half of nodes communicate vertically, and vice versa, as shown in Figure 4, allows to use only 2^{d_2-3} wavelengths in the last phase of the algorithm. An appropriate wavelength assignment can be easily obtained in a greedy fashion. So, in the special case of a square torus of $2^{d_1} \times 2^{d_1}$ nodes, only 2^{d_1-3} are thus sufficient for the dimensional exchange.

Thus, the all-to-all broadcast algorithm uses 2^{d-2} wavelengths for a ring of 2^d nodes, $2^{\max(d_1,d_2)-2}$ wavelengths for a torus of $2^{d_1} \times 2^{d_2} = 2^d$ nodes (only 2^{d_1-3} if $d_1 = d_2$), and has cost

$$\sum_{i=1}^{d} \left(\alpha + \frac{L\tau}{2^{d+1-i}} \right) = d\alpha + \left(1 - \frac{1}{2^d} \right) L\tau \qquad (3)$$

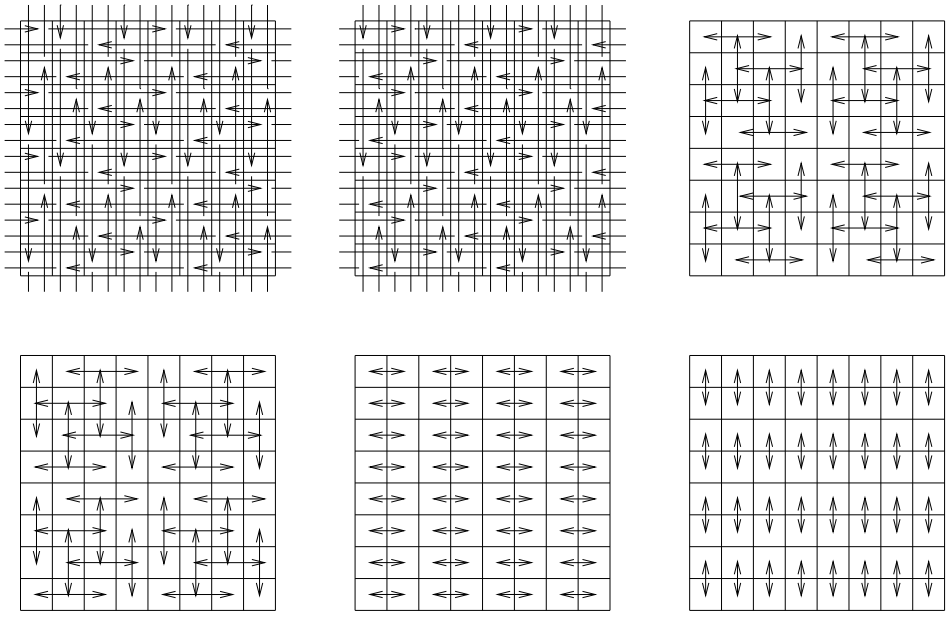

Figure 4: All-to-All broadcast algorithm using one wavelength in a torus of $2^3 \times 2^3$ nodes.

With the basic tools introduced above, we can now continue with the construction and analysis of broadcast algorithms.

3.2 Distribute-and-Exchange (DE) broadcast algorithms

The ST broadcast algorithm can be generalized to take better advantage of the available network optical bandwidth. In particular, broadcast algorithms that have bidirectional message flow are now described. They exploit the network property that messages moving in opposite directions do not contend with each other for optical communication links.

In this section we describe a family $\{\mathcal{A}(r)\}_{1\leq r\leq d}$ of broadcast algorithms, so that $\mathcal{A}(r)$ completes in $d+r$ communication rounds in a ring or a torus of 2^d nodes. For rings, our algorithms are very similar to those given in [8] for linear arrays under a circuit-switched communication model that leads nearly to the same constraints as our WDM assumptions.

3.2.1 DE broadcast algorithms on a ring

For $1 \leq r \leq d$, the ring can be viewed as 2^r interleaved subrings each with 2^{d-r} nodes. The i^{th} of these subrings, for $0 \leq i < 2^r$, consists of nodes numbered $j.2^r + i$, for $0 \leq j < 2^{d-r}$.

To broadcast the message on these interleaved rings, the message is first distributed among nodes 0, 1, ..., $(2^r - 1)$ with one wavelength. From (2), the cost of this distribution

is

$$r\alpha + \left(1 - \frac{1}{2^r}\right) L\tau \tag{4}$$

Each of the informed nodes then acts as the source of a ST broadcast of a message of length $L/2^r$ on a subring of 2^{d-r} nodes (see Figure 5). By making the even numbered nodes transmitting to the right and the odd numbered nodes to the left, only 2^{r-1} wavelengths are sufficient. From (1), the cost of this phase is

$$(d-r)\left(\alpha + \frac{L\tau}{2^r}\right) \tag{5}$$

The final phase of the algorithm consists of collecting the packets to reconstruct the original message in each node. This is accomplished by a dimensional exchange on each contiguous subring of 2^r nodes using 2^{r-2} wavelengths. We use here the fact that at the end of the ST broadcasts the distribution of the packets is the same in each contiguous subring. Therefore, receiving the information of a node in the same subring is equivalent to receiving the information of the corresponding node in an adjacent subring. From (3), this phase has cost

$$r\alpha + \left(1 - \frac{1}{2^r}\right) L\tau \tag{6}$$

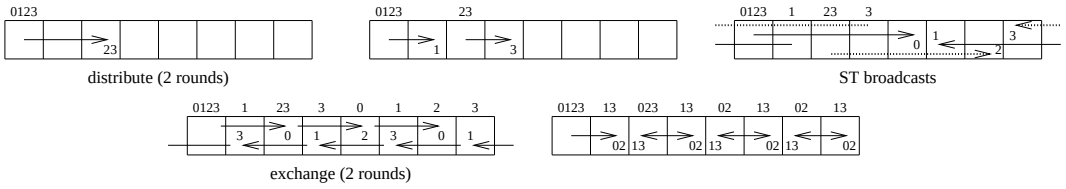

Figure 5: DE broadcast algorithm using two wavelengths in a ring of 2^3 nodes.

Thus, the DE broadcast algorithm $\mathcal{A}(r)$ in a ring of 2^d nodes uses 2^{r-1} wavelengths and its total cost is

$$(d+r)\alpha + \left(2 + \frac{d-r-2}{2^r}\right) L\tau$$

3.2.2 DE broadcast algorithms on a torus

We describe here a family of broadcast algorithms similar to those of the previous section. They complete with the same total cost but they necessitate a smaller number of wavelengths, by taking advantage of the connectivity of the torus. As before, for $1 \leq r \leq d$ the algorithm $\mathcal{A}(r)$ uses $d+r$ rounds in a torus of $2^{d_1} \times 2^{d_2} = 2^d$ nodes and proceeds in three phases.

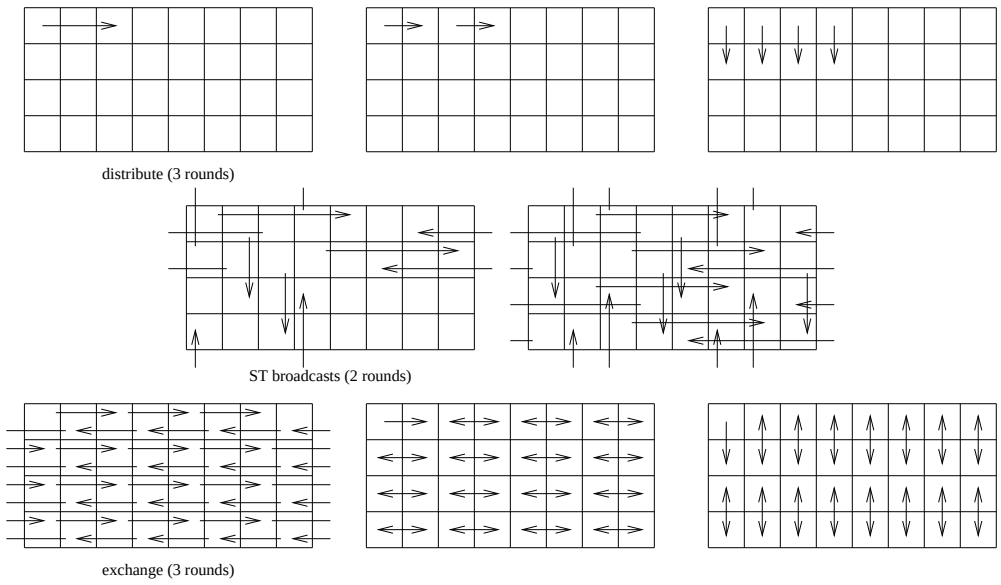

Figure 6: DE broadcast algorithm using one wavelength in a torus of $2^3 \times 2^2$ nodes.

Consider the $2^{d_1} \times 2^{d_2}$ torus as made up of 2^r interleaved subtori, each of size $2^{d_1-r_1} \times 2^{d_2-r_2}$, with $r_1 + r_2 = r$, so that each $2^{r_1} \times 2^{r_2}$ contiguous block of nodes has exactly one node from each subtorus. To have each subtorus as "square" as possible, we will choose to set $r_1 = \min(r, \lceil \frac{r-d_1+d_2}{2} \rceil)$ and $r_2 = \max(0, \lfloor \frac{r-d_1+d_2}{2} \rfloor)$ (hence $r_1 \geq r_2$), so that the difference $(d_1 - r_1) - (d_2 - r_2)$ is as small as possible.

In the first phase of the algorithm, the originator distributes the message as 2^r packets among the nodes in the $2^{r_1} \times 2^{r_2}$ block which it belongs to. See Figure 6 for an example. This necessitates only one wavelength and the same cost (4) as in a ring.

Each of the nodes in that block then acts as the root of a ST broadcast of a message of length $L/2^r$ in a subtorus of size $2^{d_1-r_1} \times 2^{d_2-r_2}$. If $(d_1 - r_1) > (d_2 - r_2)$, this phase can be realized with 2^{r_1-1} wavelengths. If $(d_1 - r_1) = (d_2 - r_2)$, alternating the orientation of the spanning trees of the 2^r subtori as in Section 3.1.3 (i.e. making one quarter of nodes communicate to the right, another downwards, another to the left and the last upwards) allows to use only 2^{r_1-2} wavelengths (See Figure 6). From Equation (1), this second phase has the same cost (5) as in a ring.

After the broadcast phase is completed, each node in each contiguous $2^{r_1} \times 2^{r_2}$ block contains one of the 2^r packets. These packets are then recombined using a dimensional exchange. As noticed before, each block can be seen as a torus of $2^{r_1} \times 2^{r_2} = 2^r$ nodes and 2^{r_1-2} wavelengths (only 2^{r_1-3} if $r_1 = r_2$) are sufficient. The last phase has the same cost (6) as in a ring.

Thus, the DE broadcast algorithm $\mathcal{A}(r)$ in a torus of $2^{d_1} \times 2^{d_2} = 2^d$ nodes uses 2^{r_1-1} wavelengths (only 2^{r_1-2} if $(r - |d_1 - d_2|)$ is non negative and even), with $r_1 = \min(r, \lceil \frac{r-d_1+d_2}{2} \rceil)$, and its total cost is

$$(d+r)\alpha + \left(2 + \frac{d-r-2}{2^r}\right) L\tau$$

3.3 Pipelined broadcast algorithms

We present in this section a generalization to rings and tori of the pipelined broadcast algorithms given in [5] for hypercubes. In addition, we provide a slight improvement on the total cost which may be also applied on hypercubes.

As before, we define a family $\{\mathcal{A}(r)\}_{r \geq d}$ of pipelined broadcast (PB) broadcast algorithms, so that $\mathcal{A}(r)$ completes in $d + r$ communication rounds in a ring or a torus of 2^d nodes. The communication pattern of all our PB algorithms are based on an embedding in rings or tori of the arc-disjoint spanning trees described in [5] for hypercubes.

However, we prefer for simplicity to describe our PB algorithms as induced by the communication pattern of the dimensional exchanges of Section 3.1.3. In consequence, the same number of wavelengths will be sufficient to ensure no wavelength conflict. It will be made use of the following property : the all-to-all broadcast is accomplished, however is chosen the order of the dimensional exchanges. This fact is easy to understand for hypercubes, where all dimensions are somewhat equivalent, and the same holds for rings and tori.

For $r \geq d$, the PB algorithm $\mathcal{A}(r)$ proceeds as follows (see Figure 7) for a ring and Figure 8 for a torus). The original message is first partitioned into $r+1$ packets, all of size $L/(r+1)$. During the first d rounds, the communication pattern of the ST broadcast is used, but the originator sends a different packet (to a different node) at each round. Each such packet is then broadcast from its first destination to all the network using a spanning tree induced by the all-to-all broadcast algorithm. During all the remaining rounds, the communication pattern of the dimensional exchanges is used cyclically. The originator continues sending a different packet at each round until round r and sends then the last packet during the last d rounds.

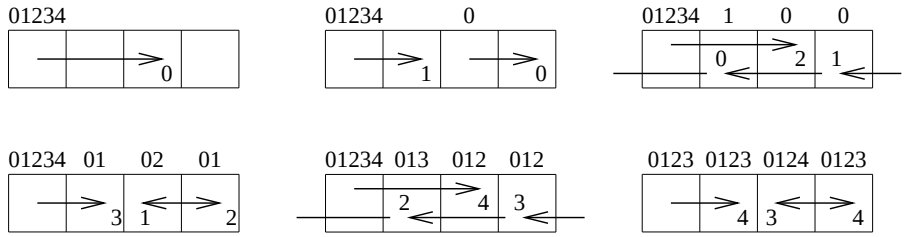

Figure 7: Pipelined broadcast algorithm in a ring of 2^2 nodes.

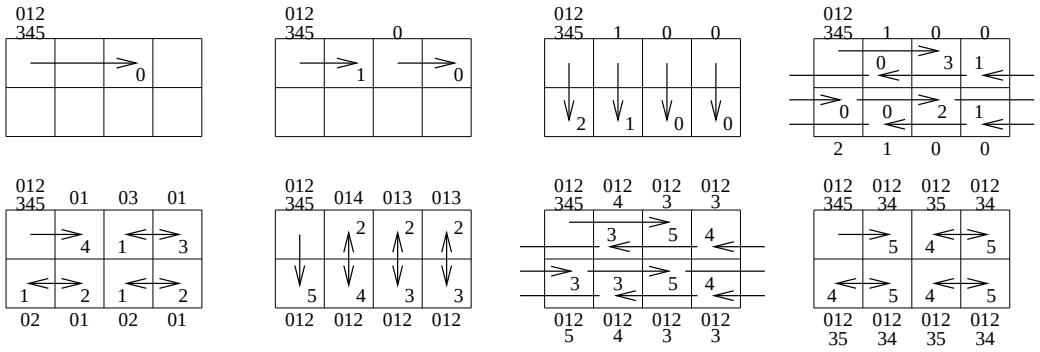

Figure 8: Pipelined broadcast algorithm in a torus of $2^2 \times 2^1$ nodes.

Thus, as packets of the same size $L/(r+1)$ are sent at any round, the PB algorithm $\mathcal{A}(r)$ has a total cost of

$$(d+r)\left(\alpha + \frac{L\tau}{r+1}\right)$$

For comparison, the algorithm of [5] completing in $(d+r)$ communication rounds has cost $(d+r)(\alpha + \frac{L\tau}{r})$. The slight improvement arises from the fact that in our algorithm the originator sends packets until the last round.

4 Result analysis

The following table summarizes the costs of the broadcast algorithms described in this paper and the maximum number of wavelengths that they use during a round. The costs of the algorithms designed for rings are given for $n = 2^d$ nodes while the torus algorithms are given for $n = 2^{d_1} \times 2^{d_2}$ nodes. It is assumed that the DE broadcast algorithms are defined for $1 \leq r \leq d$ and the PB algorithms for $d \leq r$.

Algorithm	Cost	Number of wavelengths
ring/torus ST	$d(\alpha + L\tau)$	1
ring DE	$(d+r)\alpha + \left(2 + \frac{d-r-2}{2^r}\right) L\tau$	2^{r-1}
torus DE		2^{r_1-1} if $(r - \lvert d_1 - d_2 \rvert)$ is negative or odd 2^{r_1-2} else, with $r_1 = \min\left(r, \lceil \frac{r-\lvert d_1-d_2 \rvert}{2} \rceil\right)$
square torus DE		$2^{\lfloor r/2 \rfloor}$ if r is odd $2^{r/2-2}$ if r is even
ring PB	$(d+r)\left(\alpha + \frac{L\tau}{r+1}\right)$	2^{d-2}
torus PB		$2^{\max(d_1,d_2)-2}$
square torus PB		$2^{d/2-2}$

Note that for $r \geq (\log_2 d + \Theta(1))$ the DE broadcast algorithms are not worthwhile. Indeed, their transmission cost has a constant coefficient close to two whenever $2^r = \Omega(d)$, and this cannot be improved unless $r \geq d - 1$.

For extremely short messages the ST broadcast algorithms have the lowest cost on both rings and tori because of their low latency. In fact, the ST broadcast is a DE broadcast for $r = 0$. However, the total cost function being convex, the best DE broadcast has lower cost than the ST broadcast if it is already the case for $r = 1$, that is, whenever

$$L > \frac{2\alpha}{(d-1)\tau}$$

For example, even in a 100Mb/sec network of $2^2 = 4$ nodes with latency $\alpha = 50\mu$sec, the ST broadcast has greater cost for broadcasting messages of length $L > 10$Kb.

In the case where sufficiently large messages are concerned, the PB algorithms presented in the previous section allow to decrease the coefficient of the transmission cost down to one, which is optimal. These algorithms are particularly interesting for applications such as broadcast TV or video conferencing.

In view of the results obtained in [2], all of our broadcast algorithms have nearly optimal transmission costs, and thus the best of them has the same optimality for the broadcast problem. More precisely, under the same communication model (actually not necessarily optical) it is shown in [2] that any broadcast algorithm completing in $d + r$ rounds has a transmission cost at least $(d/2^r + \Theta(1))L\tau$ if $1 \leq r \leq d$ and at least $(1 + o(1))(\frac{d+r}{r})L\tau$ if $r \geq d$.

In this paper we have not dealt with the issue of permutation costs that can affect our algorithms. As it can be noticed, internal data movements are required in the PB algorithms described in the previous section. Indeed, some of the nodes in the network receive, say, the packet numbered i up to d rounds after having received the packet numbered $(i + d - 1)$. A parameter representing the cost of moving data within a node should be subsequently included in the cost analysis of these algorithms. For our DE broadcast algorithms, this problem can be solved by modifying slightly the exchange phase. If the dimensional exchanges are realized in the reverse order of the one described in Section 3.1.3, then all the received packets can be concatenated within a node without permutation.

Nevertheless, we have chosen to present all the dimensional exchanges of our DE broadcast algorithms with that particular order for the following reason. If not enough wavelengths are available in the optical network, these algorithms can still be executed by dividing the not possible communication rounds into several rounds. For example, if one given round in the algorithm requires twice more wavelengths than it is available, then the same communication pattern can be realized in two rounds instead of one. Thus, it is important to have the communication cost of such rounds as small as possible, which is provided by our ordering of exchanges. Note that in our DE broadcast algorithms, only a few rounds require many wavelengths, therefore their total cost will not be significantly affected in case of wavelength limitation.

5 Conclusions

This study of the problem of broadcasting in WDM optical rings and two-dimensional toroidal meshes has yielded two families of algorithms. The well-known spanning tree (ST) broadcast algorithm was generalized by taking advantage of the increased bandwidth offered by the bidirectional WDM concurrent transmission. This gave rise first to a family of Distribute-and-Exchange (DE) algorithms having lower cost than the ST broadcast algorithm for all but the very short messages. Then was described a family of pipelined broadcast (PB) algorithms which perform better for long messages.

The additional connectivity of tori over rings has been fully used. Indeed, provided that the torus is quite square, the number of wavelengths used is approximatively the square root of that used in the ring with as many nodes.

Our algorithms require knowledge of machine-dependent constants for network latency and bandwidth to obtain good performance. However, implementations and performance measurements to be compared with the analytic predicted costs are desirable to design a more precise communication model. Further work should also generalize our algorithms to allow the number of nodes in a ring or in a dimension of a torus to be any number, not just a power of two.

References

[1] M. Barnett, D. G. Payne, and R. van de Geijn. Optimal broadcasting in mesh-connected architectures. Technical Report TR-91-38, Dept. of Computer Sciences, Univ. of Texas, December 1991.

[2] B. Beauquier, O. Delmas, and S. Pérennes. Tight bounds for broadcasting in the linear cost model. Technical report, Institut National de Recherche en Informatique et en Automatique, 1998. To appear.

[3] N. K. Cheung, Nosu K., and G. Winzer. Special issue on dense WDM networks. *Journal on Selected Areas in Communications*, 8, 1990.

[4] P. E. Green. *Fiber–Optic Communication Networks.* Prentice–Hall, 1993.

[5] C.T. Ho and S.L. Johnsson. Optimum broadcasting and personalized communication in hypercubes. *IEEE Transactions on Computers*, 38(9):1249–1268, September 1989.

[6] D. Minoli. *Telecommunications Technology Handbook.* Artech House, 1991.

[7] B. Mukherjee. WDM-based local lightwave networks, Part II: Multihop systems. *IEEE Network Magazine*, 6(4):20–32, July 1992.

[8] S. R. Seidel. Broadcasting on linear arrays and meshes. Technical Report ORNL/TM-12356, Oak Ridge National Laboratory, March 1993.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, B.P. 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399