Time-Optimal Gossip in Noncombining 2-D Tori with Constant Buffers*

Michal Šoch and Pavel Tvrdík

Department of Computer Science and Engineering
Czech Technical University, Karlovo nám. 13
121 35 Prague, Czech Republic
{soch,tvrdik}@sun.felk.cvut.cz
http://cs.felk.cvut.cz/pcg/

Abstract. This paper describes a time-, transmission-, and memory-optimal algorithm for gossiping in general 2-D tori in the noncombining full-duplex and all-port communication model.

1 Introduction

Gossiping, also called *all-to-all broadcasting*, is a fundamental collective communication problem: each node of a network holds a packet that must be delivered to every other node and all nodes start simultaneously. In this paper, we consider gossiping in the *noncombining store-and-forward all-port full-duplex* model. In such a model, a gossip protocol consists of a sequence of *rounds* and broadcast trees can be viewed as *levels* of arc sets, one level for one round. Given a broadcast tree $BT(u)$ rooted in vertex u of a graph G, the arcs crossed by packets in round t are denoted by $\mathcal{A}_t(BT(u))$. The *height* of a broadcast tree $BT(u)$, denoted by $h(BT(u))$, is the number of levels of BT, i.e., the number of rounds of a broadcast using this tree. Given $BT(u)$ and $i < h(BT(u))$, the i-th *level subtree* of $BT(u)$, denoted by $BT^{[i]}(u)$, is the subtree of $BT(u)$ consisting of the first i levels.

Given a graph G, every node of G has to receive $|\mathcal{V}(G)| - 1$ packets and in one round, it can receive $\delta(G)$ packets in the worst case, where $|\mathcal{V}(G)|$ is the size of G and $\delta(G)$ is the minimum degree of a node in G. The lower bound on the number of rounds of a gossip in G is therefore $\tau_g(G) = \left\lceil \frac{|\mathcal{V}(G)|-1}{\delta(G)} \right\rceil$.

If G is a vertex-transitive network, the gossip problem can be solved by constructing a *generic* broadcast tree, denoted by $BT(*)$, whose root can be placed into any vertex of G. All broadcast trees are therefore isomorphic.

2-D torus $T(m,n)$ is a cross product of 2 cycles of size m and n. Vertex set of $T(m,n)$ is the cartesian product $\{0,..,m-1\} \times \{0,..,n-1\}$.

Tori are vertex transitive and the isomorphic copies of the generic broadcast trees are made by *translation*.

* This research was supported by GAČR Agency under Grant 102/97/1055, FRVŠ Agency under Grant 1251/98, and CTU Grant 3098102336.

Definition 1. *Given nodes* $u = [u_x, u_y]$ *and* $v = [v_x, v_y]$ *of* $T(m,n)$, *a* translation from u to v, *denoted by* $\psi_{u \to v}$, *is induced by node mapping* $([w_x, w_y] \mapsto [w_x \oplus_m v_x \ominus_m u_x, w_y \oplus_n v_y \ominus_n u_y]$, *where the addition and subtraction is taken modulo* m *and* n, *respectively.*

A gossip in $T(m,n)$ is time- and transmission-optimal iff $h(BT(*)) = \tau_g(T(m,n)) = \lceil \frac{mn-1}{4} \rceil$ and all isomorphic copies of $BT(*)$ are pairwise *time-arc-disjoint*, i.e., two arcs at the same time-level of any two broadcast tree are never mapped on the same arc of $T(m,n)$ and the communication in every round is therefore contention-free. It is known that in tori, time-arc-disjointedness is equivalent to the distinctness of directions of arcs at every level of the generic tree. The set of directions of arcs $\mathcal{A}_t(BT(*))$ is denoted by $\mathrm{dir}(\mathcal{A}_t(BT(*)))$. In a 2-D torus there exist four directions, usually denoted by N,E,W,S. N-S direction is vertical, W-E direction is horizontal. Hence, a sufficient condition for $BT(*)$ to guarantee a time-optimal gossip is that for any $t < \tau_g(T(m,n))$, every $\mathcal{A}_t(BT(*))$ is a set of 4 arcs of 4 distinct directions N, E, W, S.

In general, a gossip algorithm in $T(m,n)$ may require additional buffers in routers for packets which must wait at least one round before they can be sent out and the routers can get rid of them. Let $\beta(G)$ denote the maximum size of auxiliary buffers per router during gossiping in network G.

In [2], we have presented a time-optimal gossip protocol for general $T(m,n)$. However, this algorithm is not memory-optimal, it requires auxiliary buffers for $\Theta(\max(n, m-n))$ packets per router. The generic broadcast tree of $T(m,n)$ in [2] is built in two steps: filling up the maximal square submesh of odd side + informing the rest of nodes. Additional buffers for packets are required on the sides of the square submesh.

In this paper, we present a time- and memory-optimal gossip algorithm for $T(m,n)$ which requires $\beta(T(m,n)) = 3$. To keep the size of auxiliary buffers constant, the generic broadcast tree is built from vertical stripes of width 2. Only constant number of packets must be stored for dissemination in directions W and E, which allows to concatenate vertical stripes horizontally.

2 Generic Broadcast Trees for Optimal Gossip on $T(m,n)$.

Theorem 1. *For any* $m \geq n \geq 2$, *there exists a generic* $BT(*)$ *of* $T(m,n)$ *such that* $h(BT(*)) = \lceil \frac{mn-1}{4} \rceil$ *and* $|\mathcal{A}_t(BT(*))| = |\mathrm{dir}(\mathcal{A}_t(BT(*)))|$ *for all* $1 \leq t \leq h(BT(*))$ *and* $\beta(T(m,n)) \leq 3$. *If* $n = 2$ *and* m *is even, one extra round is needed.*

Proof. For $m = n$, the algorithm is trivial and $\beta(T(m,m)) = 0$ if m is odd and $\beta(T(m,m)) = 1$ otherwise. Assume without losing generality $m > n$. The construction of time-arc-disjoint trees for time- and memory-optimal gossip depends slightly on the values of m and n, the number of different cases is 7, the basic idea is, however, the same in all of them, see [3] for details.

The memory requirements of our algorithm are stated in the following table.

	n=2	n=3, $m \leq 5$	n odd	$n \geq 4$ even, m odd	$n \geq 4$ even, m even
$\beta(T(m,n))$	0	0	2	2	3

In this paper, we describe only one particular case when $m > n \geq 5$ are odd numbers. To make the construction of $BT(*)$ in $T(m,n)$ as simple as possible, the same patterns of arc sets $\mathcal{A}_t(BT(*))$ are used repeatedly in various rounds t. The whole generic tree $BT(*)$ is then built using several arc patterns. Arc pattern i is depicted on Figures 1 and 2 as a quadruple of arcs labeled i. We associate with every pattern i a so called *expansion operator*, denoted by Γ_i. The broadcast tree is then specified by a regular expression over expansion operators. If $\mathcal{A}_t(BT(*))$ has pattern i, an *attachment* of arcs at level t to $(t-1)$-level generic subtree is described as an *application* of the corresponding operator $BT^{[t]}(*) = BT^{[t-1]}(*)\Gamma_i$.

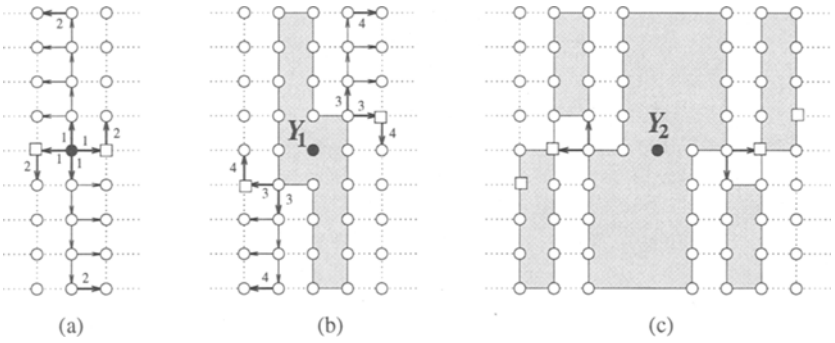

Fig. 1. The first phase of building $BT(*)$ in $T(m,n)$, if $m \geq 11$, $m > n \geq 5$ are odd. (a) $Y_1 = \Gamma_1^{\frac{n-1}{2}}\Gamma_2$. (b) $Y_2 = Y_1\Gamma_3^{\frac{n-3}{2}}\Gamma_4$. (c) Y_2^2.

The generic tree is built in two phases. Figure 1(a) depicts the first phase of constructing $BT(*)$ if $m \geq 11$. If $m \leq 9$, the first phase is void. The black circle is the root of $BT(*)$. The square symbols $\square$ denoted the nodes which have to store packets needed in later steps. Repeated application of Γ_1 in Figure 1(a) is followed by Γ_2 which produces $\frac{n+1}{2}$ level subtree $Y_1 = BT^{[\frac{n+1}{2}]}(*) = \Gamma_1^{\frac{n-1}{2}}\Gamma_2$. Note that the expansion by Γ_1 proceeds in N-S direction. Further $\frac{n-3}{2}$ levels are attached in direction N-S by repeated application of Γ_3 and after applying Γ_4, we get n-level subtree $Y_2 = BT^{[n]}(*) = Y_1\Gamma_3^{\frac{n-3}{2}}\Gamma_4$. If $m \geq 15$, the whole process can be repeated by expanding Y_2 in W-E direction, see Figure 1(c). The second phase depends on $m \bmod 4$. Let us describe the case of $m \equiv 1 \bmod 4$ (the other case is similar). Let $Y_3 = Y_2^{\frac{m-9}{4}}$. If $m = 9$, then Y_3 shrinks to the root.

Figure 2 depicts the solution. We interpret the use of expansion operators similarly as in the first phase. In $\frac{3n-1}{2}$ rounds, we expand Y_3 to $Y_4 = Y_3Y_1\Gamma_3^{\frac{n-3}{2}}\Gamma_5\Gamma_3^{\frac{n-3}{2}}\Gamma_4$ (see Figure 2(a)). Y_4 is then expanded in $3\lfloor\frac{n-1}{4}\rfloor$ rounds using Γ_6 and Γ_7 (see Figure 2(b)). Pattern $\Gamma_6^2\Gamma_7$ diffuses vertically until the

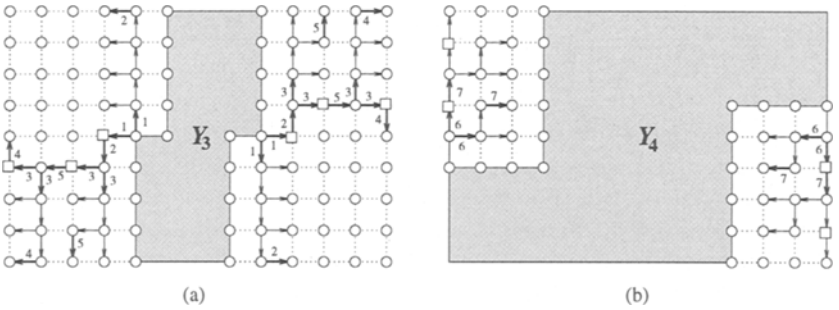

Fig. 2. The second phase of constructing $BT(*)$, if $m > n \geq 5$ are odd, $m \equiv 1 \bmod 4$. (a) Construction of Y_4. (b) Application of Γ_6 and Γ_7 patterns.

boundary is reached. Finally, if $n \equiv 1 \bmod 4$ (as shown on Figure 2(b)), one final round is needed, and if $n \equiv 3 \bmod 4$, after two more applications of Γ_6, only two uninformed nodes surrounded by informed ones remain.

The memory requirements for this case of m and n follow easily. In the first phase, 2 packets must be stored to apply Γ_2 (see the square symbols in Figure 1(a)). Then, these buffers can be reused to store packets for application of Γ_4 (see Figure 1(b)). The second phase is similar. In every round of the gossip, no more than two packets must be stored at a time. □

3 Conclusions

The minimal-height time-arc-disjoint trees in the proof of Theorem 1 provide time-, transmission-, and memory-optimal gossip algorithm in noncombining full-duplex all-port 2-D tori. The algorithm requires routers with buffers for at most 3 packets. An interesting problem is to find the exact lower bound on the size of additional buffers. Another open problem is to find a time- and memory-optimal gossip algorithm for 2-D meshes.

References

1. J.-C. Bermond, T. Kodate, and S. Perennes. Gossiping in Cayley graphs by packets. In M. Deza, et. al., editors, *Combinatorics and Computer Science*, LNCS 1120, pages 301–315. Springer, 1995.
2. M. Šoch and P. Tvrdík. Optimal gossip in store-and-forward noncombining 2-D tori. In C. Lengauer, et. al., editors, *Euro-Par'97 Parallel Processing*, LNCS 1300, pages 234–241. Springer, 1997. Research report at `http://cs.felk.cvut.cz/pcg/`.
3. M. Šoch and P. Tvrdík. Time-optimal gossip in noncombining 2-D tori with constant buffers. Manuscript at `http://cs.felk.cvut.cz/pcg/`.