アルゴリズム 36－6
（1993．11．25）

長さの制限されたパケットによるメッシュバス計算機上のゴシッピング

藤田　聡

広島大学工学部第二類

本稿では二次元状に配置されたn^2個のノード(処理要素)とそれらを結ぶn本の行バスとn本の列バスとからなるメッシュバス計算機上でのゴシッピング問題を考える。ゴシッピング問題とは、各ノードがそれぞれ初期状態で保持している固有のトークンを系のすべてのノード間で互いに交換する問題である。ここでは各バスはCREWでアクセスされるものとし、各メッセージは高々ℓ個のトークンを運ぶことができるものと仮定する。

本稿では、*SIMPLE*, *PARTITION*, *CENTRALIZE* の3つのアルゴリズムを提案する。各アルゴリズムは、それぞれある条件を満たすℓに対して、漸近的に最適な実行時間でゴシップを完了する。

結果として、長さが$\Theta(n)$以外のすべてのℓに対して漸近的に最適な実行時間でのゴシッピングが行えることが示される。

Keywords: メッシュバス計算機, ゴシップ問題, パケット通信

Gossiping in Mesh-Bus Computers by Packets with Bounded Length

Satoshi Fujita

Department of Electrical Engineering
Faculty of Engineering, Hiroshima University
Email: fujita@csl.hiroshima-u.ac.jp

This report considers the gossiping problem in mesh-bus computers, in which n^2 nodes (i.e., processing elements) arranged on $n \times n$ array are connected by n column-buses and n row-buses. Let V be the set of nodes. The gossiping is a problem of exchanging $|V|$ tokens, each of which is initially held by distinct node, among all nodes in V. In this report, we assume that each shared bus is accessed in CREW manner, and that each message can carry at most ℓ tokens in each step. This report proposes three algorithms; *SIMPLE*, *PARTITION*, and *CENTRALIZE*, each of which asymptotically achieves a lower bound on the gossiping time for a range of ℓ.

Keywords: mesh-bus computer, gossiping, bounded packet length

1 Introduction

A mesh-bus computer is a parallel processor in which nodes (i.e., processing elements) are arranged on a two-dimensional array, and nodes on each row and nodes on each column, respectively, are connected by a shared bus. Each shared bus is accessed in CREW manner, i.e., all nodes connected with a bus can eavesdrop on the bus at the same time, while in each time, at most one node is allowed to send a message through the bus. So far, there have been proposed several parallel algorithms for mesh-bus computers [1, 2, 5, 6, 9, 10, 11].

Suppose that at the initial state, each node contains a piece of information called a *token*. The gossiping problem, which has been investigated extensively during the last decade [4, 7, 8], is a problem of exchanging tokens among all nodes in the system. This report addresses the gossiping on mesh-bus computers. In [3], we have proposed a gossiping algorithm for $n \times n$ mesh-bus computer, which takes $\lfloor n/2 \rfloor + \lceil \log_2 n \rceil + 2$ steps provided each message can carry *any* number of tokens in a step[1]. This report will extend the result. In the following, we consider the same problem under the following (more realistic) assumption on the communication: *Each message can carry at most* ℓ $(1 \leq \ell \leq n^2)$ *tokens in a step.* A key to construct efficient algorithm under the assumption is to increase the throughput of the communication, e.g., (1) by packing as many tokens as possible $(\leq \ell)$ into each message, and/or (2) by utilizing almost all buses during the gossiping process.

This report is organized as follows. Section 2 provides some definitions. Section 3 considers lower bounds on the gossiping time. Sections 4, 5, and 6 propose gossiping algorithms for $\ell = o(n)$, $\ell = \omega(n)$, and for $\ell = \Theta(n)$, respectively.

[1] It is at most only 3 more steps than a lower bound [3] (c.f. Theorem 1).

2 Preliminaries

A *mesh-bus computer* $\mathcal{M}$ consists of n^2 nodes arranged on $n \times n$ array. Denote the node located at the ith row and the jth column by (i, j). Let $V = N \times N$ be the set of nodes in $\mathcal{M}$, where $N = \{1, 2, \ldots, n\}$. The nodes in $\mathcal{M}$ are connected with n row-buses and n column-buses. Node $(i, j) \in V$ is connected with row-bus R_i and column-bus C_j. $\mathcal{M}$ has a global clock, and all nodes execute their operations synchronously according to the global clock. We call a unit time of the clock a *step*. Each node (i, j) always eavesdrops on two buses R_i and C_j to receive all tokens flowing on those buses, and can send a *message* containing tokens through the two buses, while in each step, at most one node is allowed to send a message through a bus (i.e., each bus is accessed in CREW manner).

Each node (i, j) in $\mathcal{M}$ initially holds a token denoted by $t(i, j)$, and we assume that each message can carry *at most* ℓ tokens. The *gossiping problem* we consider in this report is described as follows: *For each* $(i, j) \in V$, *broadcast* $t(i, j)$ *to all nodes in* V. For simplicity, we suppose that all broadcast requests are issued simultaneously.

3 Lower Bounds

For sufficiently large ℓ, we have obtained a lower bound on the gossiping time [3].

Theorem 1 *When* $\ell = \infty$, *i.e., when each message can carry any number of tokens, the gossiping requires at least* $\lceil n/2 \rceil + \lceil \log_2 n \rceil - 1$ *steps.* □

On the other hand, it needs at least $n + 1$ distinct buses to broadcast a token to all nodes in V. Hence during a gossiping process, $n^2 \times (n + 1)$ copies of tokens have to be carried by $2n$ buses. Since each bus can carry at

most ℓ tokens in a step, we have another lower bound as follows:

Theorem 2 *When each message can carry at most ℓ tokens, it requires at least $n^2/2\ell + n/2\ell$ steps to complete the gossiping.* □

4 Short Packets

This section considers the case of $\ell = o(n)$; i.e., proposes an efficient algorithm for packets with short length.

4.1 Basic Algorithm

First, consider the following algorithm.

Algorithm *SIMPLE*

Phase 1: If $i + j$ is even (resp. odd), then node (i, j) sends $t(i, j)$ through row-bus R_i (resp. column-bus C_j).

Phase 2: Let $H(i, j)$ (resp. $V(i, j)$) be the set of tokens received by node (i, j) through row-bus R_i (resp. column-bus C_j) in Phase 1. Each node $(i, j) \in V$ sends $V(i, j)$ (resp. $H(i, j)$) through row-bus R_i (resp. column-bus C_j). □

See Figure 1 for illustration. The correctness of the algorithm is clear. Let $\tau = \lceil n/2 \rceil$. Phase 1 takes τ steps, since each bus is sequentially accessed by at most τ nodes. On the other hand, Phase 2 takes

$$\begin{aligned} n\left\lceil \frac{\tau}{\ell} \right\rceil &\leq n\left\{ \left(\frac{n+1}{2}\right)\left(\frac{1}{\ell}\right) + 1 - \frac{1}{\ell} \right\} \\ &= \frac{n^2}{2\ell} - \frac{n}{2\ell} + n \end{aligned}$$

steps. Hence we have the following theorem.

Theorem 3 *Algorithm SIMPLE completes the gossiping correctly on $\mathcal{M}$ in at most $n^2/2\ell - n/2\ell + \lceil 3n/2 \rceil$ steps.* □

Corollary 1 *When $\ell = o(n)$, algorithm SIMPLE asymptotically achieves a lower bound on the gossiping time on $\mathcal{M}$.* □

4.2 Refinement

If ℓ divides τ $(= \lceil n/2 \rceil)$, we have the following corollary of Theorem 3.

Corollary 2 *If ℓ divides $\lceil n/2 \rceil$, algorithm SIMPLE completes the gossiping on $\mathcal{M}$ in $n^2/2\ell + n/2\ell + \lceil n/2 \rceil$ steps.* □

Note that the upper bound in the corollary is at most $\lceil n/2 \rceil$ more steps than the lower bound in Theorem 2.

On the other hand, suppose that ℓ $(= o(n))$ does not divide τ, and that $\ell < \tau$. Note that the inequality always holds for sufficiently large n, since $\ell = o(n)$. Rewrite τ as $\alpha\ell + \beta$, where $\beta = \tau \pmod{\ell}$ and $\beta \neq 0$.

In Phase 1 of algorithm *SIMPLE*, each node sends exactly one token through a bus. However, since each message can carry at most ℓ tokens, during Phase 1, each node can send at most $\ell - 1$ other tokens having been received in previous steps of Phase 1.

In the tth $(1 \leq t \leq \ell)$ step of Phase 1, node u has received $|H(u)| = |V(u)| = t - 1$ tokens from other nodes on the same row or column. Hence, the above modification reduces the running time to

$$\begin{aligned} T &\leq \sum_{i=1}^{\ell} \left\lceil \frac{\tau - (i-1)}{\ell} \right\rceil \\ &\quad + \{(n - \tau - \ell)\alpha + \tau(\alpha + 1)\} \\ &= \left\{ \sum_{i=1}^{\beta} (\alpha + 1) + \sum_{i=\beta+1}^{\ell} \alpha \right\} \\ &\quad + (n - \ell)\alpha + \tau \\ &= (\ell\alpha + \beta) + (n - \ell)\alpha + \tau \\ &= n\alpha + \beta + \tau. \end{aligned}$$

Since $\alpha = (\tau - \beta)/\ell$, we have

$$\begin{aligned} T &\leq n\left(\frac{n+1}{2} - \beta\right)\left(\frac{1}{\ell}\right) + \beta + \tau \\ &= \frac{n^2}{2\ell} + \frac{n}{2\ell} + \beta\left(1 - \frac{n}{\ell}\right) + \left\lceil \frac{n}{2} \right\rceil. \end{aligned}$$

When $\ell < n$, $\beta = 1$ gives the maximum value to T. Hence we have the following theorem.

Theorem 4 *When $\ell < \lceil n/2 \rceil$ and ℓ does not divide $\lceil n/2 \rceil$, the gossiping on $\mathcal{M}$ takes at most $n^2/2\ell - n/2\ell + n + 2$ steps.* □

It is at most $n - n/2\ell + 2$ more steps than a lower bound.

5 Long Packets

This section considers the case of $\ell = \omega(n)$. The reader can easily verify that when $\ell = \omega(n)$, algorithm *SIMPLE* is not (asymptotically) tight (it takes at least n steps, while the lower bound in Theorem 1 is $n/2 + o(n)$).

5.1 When $\ell = o(n^2)$

Suppose that $\ell = o(n^2)$. Let $m = \lfloor 2\ell/n \rfloor$ and $\sigma = \lceil n/m \rceil$ (i.e., $m = o(n)$ and $\sigma = o(n)$). We first partition N into σ subsets $N_1, N_2, \ldots, N_\sigma$ such that $N_i = \{(i-1)m + 1, \ldots, im\}$ for all $1 \leq i < \sigma$ and $N_\sigma = N \backslash \bigcup_{1 \leq i < \sigma} N_i$. Under the partition, consider the following algorithm.

Algorithm *PARTITION*

Phase 1: If $i + j$ is even (resp. odd), then node (i, j) sends token $t(i, j)$ through row-bus R_i (resp. column-bus C_j).

Phase 2: Let $H_1(u)$ (resp. $V_1(u)$) be the set of tokens received by node u through row-bus (resp. column-bus) in Phase 1. For all $1 \leq x \leq \sigma$ in parallel, each node $u \in N_x \times N_x$ sends $V_1(u)$ (resp. $H_1(u)$) through row-bus (resp. column-bus).

Phase 3: Let $H_2(u)$ (resp. $V_2(u)$) be the set of tokens received by node u through row-bus (resp. column-bus) in Phase 2. For each $i \in N$, node $u \in \{(i, i + tm) : t \in Z\} \cap V$, where Z is the set of integers, sends $V_2(u)$ through row-bus R_i. For each $j \in N$, node $w \in \{(j + tm, j) : t \in Z\} \cap V$ sends $H_2(w)$ through column-bus C_j. □

See Figure 2 for illustration.

Theorem 5 *Algorithm PARTITION completes gossiping correctly on $\mathcal{M}$.*

Proof. Let (i, j) be a node in V. Assume that $i + j$ is odd. (When $i + j$ is even, we can prove the same statement in a similar way.)

Suppose that $i \in N_x$. In Phase 1, $t(i, j)$ is broadcast to all nodes on the ith row, i.e., all nodes in $U = \{(i, y) : y \in N_x\}$. In Phase 2, node $(i, k) \in U$ sends $t(i, j)$ $(\in H_1(i, k))$ to all nodes on the kth column. In other words, Phase 2 broadcasts $t(i, j)$ to all nodes in $W = N \times N_x$. Now, for each row, there is a node in

$$\{(x, y) : y \in N_x\} \cap \{(x, i + tm) : t \in Z\},$$

which broadcasts the token to all nodes on the same row in Phase 3. Hence $t(i, j)$ is broadcast to all nodes in V. □

Theorem 6 *When $\ell = \omega(n)$ and $\ell = o(n^2)$, algorithm PARTITION requires at most $n/2 + o(n)$ steps, which asymptotically achieves a lower bound.*

Proof. Let $\tau = \lceil n/2 \rceil$. Phase 1 takes τ steps. In Phase 2, each bus is sequentially accessed by m $(= \lfloor 2\ell/n \rfloor)$ nodes, each of which sends at most τ tokens through one bus. It holds $\ell > \tau$ since $m = o(n)$ and $\ell = \omega(n)$. Hence Phase 2 takes at most m steps. In Phase 3, each bus is sequentially accessed by σ nodes, each of which sends at most $m\tau$ $(\leq \ell)$ tokens. Hence

Phase 3 takes at most σ steps. Consequently, algorithm *PARTITION* takes at most

$$\tau + m + \sigma \;=\; \frac{n}{2} + o(n)$$

steps. □

5.2 When $\ell = \Theta(n^2)$

Next consider the case of $\ell = \Theta(n^2)$. By definition, there is a constant c such that $\ell \geq cn^2$.

Let $D = \{(x,x) : x \in N\}$. In [3], we have proposed an efficient gossiping algorithm provided $\ell \geq n^2$.

Algorithm *CENTRALIZE*

Phase 1: If $i+j$ is even and $i \neq j$, send $t(i,j)$ to node (i,i) $(\in D)$ through row-bus R_i. If $i+j$ is odd, send $t(i,j)$ to node (j,j) $(\in D)$ through column-bus C_j.

Phase 2: Collect tokens held by nodes in D to a node in V as follows:

Step 1: Let $U = D$.

Step 2: Let $U = \{x_1, x_2, \ldots, x_{|U|}\}$. If $|U| = 1$, then go to Phase 3.

Step 3: Let $W = \{(x_{2i-1}, x_{2i}) : 1 \leq i \leq \lfloor |U|/2 \rfloor\}$.

Step 4: Let $U = \emptyset$.

Step 5: For each $((a,b),(c,d)) \in W$, (a,b) sends a message to (a,d) through row bus R_a; (c,d) sends a message to (a,d) through column bus C_d; and add (a,d) to U.

Step 6: Go to Step 2.

Phase 3: Let u be the node in U. Node u broadcasts the set of all tokens to all nodes in V. □

See Figure 3 for illustration.

Theorem 7 *When $\ell = \Theta(n^2)$, algorithm CENTRALIZE completes the gossiping on $\mathcal{M}$ in $n/2 + O(\log_2 n)$ steps.*

Proof. At the end of Phase 1, for any $v \in V$ $(= V_0 \cup V_1)$, $t(v)$ is held by a node in D, since there is a node in D on each row and column. Since $\lfloor n/2 \rfloor$ nodes exclusively send messages through each bus, Phase 1 takes $\lfloor n/2 \rfloor$ steps.

In Phase 2, the nodes in D collect tokens to a node in V. Clearly, if nodes in U share no buses, then nodes in new U created in Step 5 share no buses. Since each message can carry cn^2 tokens for some constant c, Phase 2 takes $O(\log_2 n)$ steps.

At the end of Phase 2, a node $u \in U$ holds the set of 'all' tokens. In Phase 3, node u broadcasts the set to all nodes in V, which takes $O(1)$ steps. □

6 Other Cases

Finally, consider the case of $\ell = \Theta(n)$. Suppose that $\ell \leq cn$ for some constant c.

When $\ell < n/2$, by Theorem 4, an improved version of algorithm *SIMPLE* in Section 4.2 completes the gossiping in at most

$$\begin{aligned} \frac{n^2}{2\ell} - \frac{n}{2\ell} + n + 2 \;&\leq\; \frac{n^2}{2\ell} + n + 2 \\ &\leq\; (1+2c)\frac{n^2}{2\ell} + 2 \end{aligned}$$

steps, where we use $1/n \leq c/\ell$. When $\ell < n/2$, since $c < 1/2$, it is at most $n/\ell + O(1)$ steps.

On the other hand, when $\ell > n/2$, we can modify algorithm *PARTITION* in such a way that $m = \lfloor 2c \rfloor$ and $\sigma = \lceil n/m \rceil$.

Note that $2\ell < mn$ (i.e., $n/m < n^2/2\ell$) since $\ell \leq cn$. By Theorem 6, the modified version of algorithm *PARTITION* finishes the gossiping in

$$\left\lceil \frac{n}{2} \right\rceil + m + \sigma \;\leq\; \frac{n}{2} + \frac{n^2}{2\ell} + O(1)$$

steps. When $\ell \leq n$, it takes at most

$$\left(1 + \frac{\ell}{n}\right)\frac{n^2}{2\ell} + O(1) \;\leq\; 2 \times \frac{n^2}{2\ell} + O(1)$$

$$\leq \quad 2 \times \frac{n}{2} + O(1)$$

steps. Hence we have the following theorem.

Theorem 8 *When $\ell = \Theta(n)$, the gossiping completes in at most $2 \times \max\{n/2, n^2/2\ell\} + O(1)$ steps, which is (asymptotically) at most twice of a lower bound.* □

Acknowledgement

The research was partly supported by the Grant in Aid for Scientific Research of the Ministry of Education, Science and Culture of Japan.

References

[1] A. Bar-Noy and D. Peleg. Square meshes are not always optimal. *IEEE Trans. Comput.*, 40(2):196–204, February 1991.

[2] Y.-C. Chen, W.-T. Chen, G.-H. Chen, and J.-P. Sheu. Designing efficient parallel algorithms on mesh-connected computers with multiple broadcasting. *IEEE Trans. Parallel and Distributed Systems*, 1(2):241–246, April 1990.

[3] S. Fujita and M. Yamashita. Optimal gossiping in mesh-bus computers. *Parallel Processing Letters*, 1993. (submitted).

[4] S. M. Hedctniemi, S. T. Hedetniemi, and A. L. Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18:319–349, 1988.

[5] K. Iwama and Y. Kambayashi. An O(log n) parallel connectivity algorithm on the mesh of buses. In *Proc. 11th IFIP World Computer Congress*, pages 305–310. IFIP, 1989.

[6] K. Iwama, E. Miyano, and Y. Kambayashi. Bounds for routing on the mesh-bus computer. manuscript, 1991.

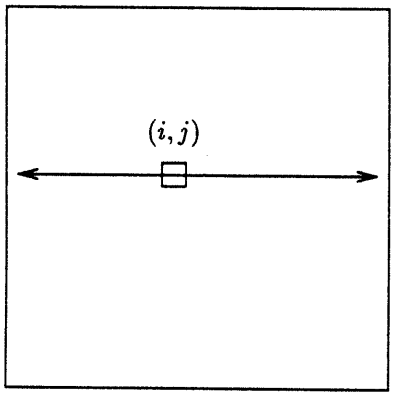
[7] D. W. Krumme. Fast gossiping for the hypercube. *SIAM J. Comput.*, 21(2):365–380, April 1992.

[8] D. W. Krumme, G. Cybenko, and K. N. Venkataraman. Gossiping in minimal time. *SIAM J. Comput.*, 21(1):111–139, February 1992.

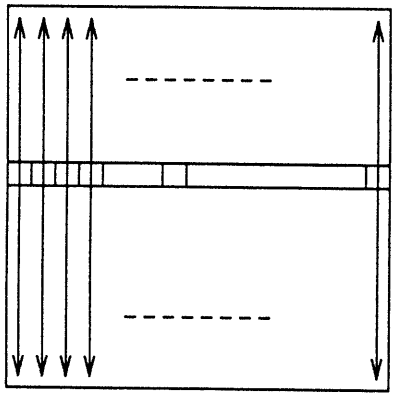
[9] V.K. P. Kumar and C. S. Raghavendra. Array processor with multiple broadcasting. In *Proc. 12th ISCA*, pages 2–10. IEEE/ACM, 1985.

[10] K. Nakano, T. Masuzawa, and N. Tokura. Optimal sorting algorithms on processor arrays with multiple buses. *Tech. Report COMP, IEICE Japan*, 91-7, 1991.

[11] Q. F. Stout. Meshes with multiple buses. In *Proc. 27th FOCS*, pages 264–273. IEEE, 1986.

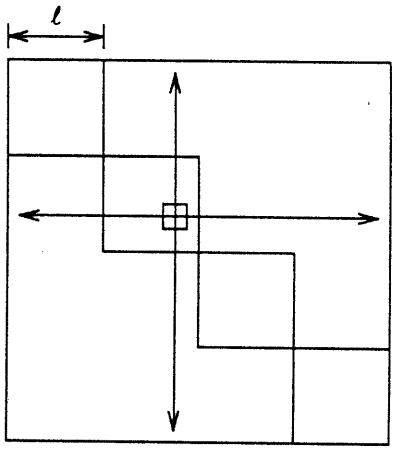


Phase 1.

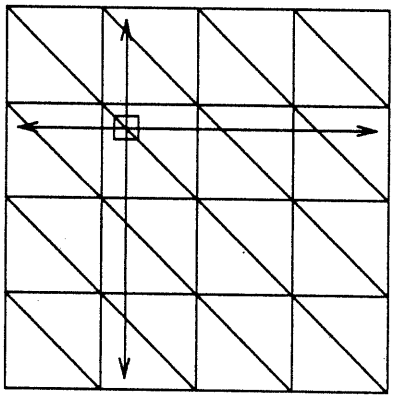


Phase 2.

Figure 1: Algorithm *SIMPLE* (when $i+j$ is even).

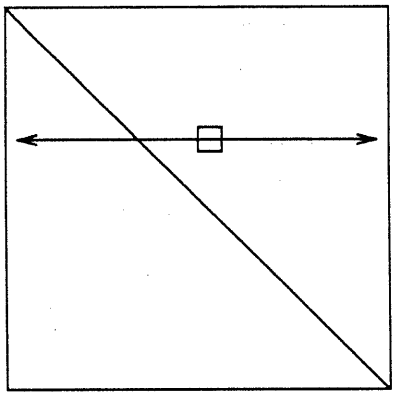


Phase 2.

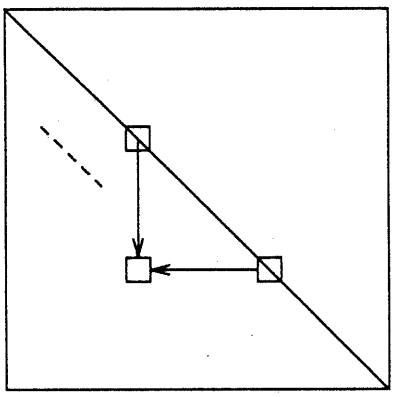


Phase 3.

Figure 2: Algorithm *PARTITION*.



Phase 1.



Phase 2.

Figure 3: Algorithm *CENTRALIZE*.