LAPACK Working Note 29: On Global Combine Operations

Robert van de Geijn

CRPC-TR91137
May, 1991

Center for Research on Parallel Computation
Rice University
P.O. Box 1892
Houston, TX 77251-1892

LAPACK Working Note 29
On Global Combine Operations *

Robert A. van de Geijn

Department of Computer Science
University of Tennessee
Knoxville, Tennessee 37996-1301

and

Department of Computer Sciences
University of Texas at Austin
Austin, Texas 78712

na.vandegeijn@na-net.ornl.gov

May 23, 1991

Abstract

We discuss a hybrid strategy for implementing global combine operations on distributed memory MIMD multicomputers. A theoretical analysis is given and results from its implementation on the Intel iPSC/860 are reported.

1 Introduction

In this paper, we address the implementation of the combine operation when vectors of data to be combined are distributed among the processors (nodes) of a MIMD hypercube architecture. Several solutions to this problem have appeared in the literature. We propose a hybrid approach that combines two of the most successful of these solutions.

Our interest in the implementation of global combine operations stems from their use in parallel implementations of numerical algorithms for dense matrix problems. For example, the global summation of vectors is useful when implementing the reduction a matrix to Hessenberg or tridiagonal form [2]. Several global combine operators are included in a proposed set of communication primitives, the Basic Linear Algebra Communication Subprograms (BLACS) [1].

*This work was supported by the National Science Foundation Science and Technology Center Cooperative Agreement No. CCR-8809615

2 Assumptions

Target architectures for our algorithm are distributed memory hypercube multicomputers, including Multiple Instruction Multiple Data (MIMD) machines like Intel's iPSC/860, NCUBE Inc.'s NCUBE2 and Transputer based machines.

For our theoretical analysis, we will assume the following model:

1. The hypercube consists of $p = 2^d$ nodes, labeled $\mathbf{P}_0, \ldots, \mathbf{P}_{p-1}$.
2. Node $\mathbf{P}_i$ has physical neighbors $\mathbf{P}_{i_k}$, $k = 0, \ldots, d-1$ where i_k equals the integer that differs from i in the kth binary digit. We will use the notation $\mathtt{ngbr}(i,k)$ to denote i_k and $\mathtt{bit}(i,k)$ to denote the kth binary digit of i. For fixed k, $\mathbf{P}_{i_k}$ is called the neighbor of $\mathbf{P}_i$ in direction k.
3. A node can exchange data with only one neighbor at a time, i.e., it can simultaneously send to and receive from the same neighbor [1], but not to and from different neighbors.
4. Communication and computation do not overlap, i.e., the communication is blocking.
5. Exchanging messages of length n between two neighbors requires time $\alpha + n\beta$, where α and β represents the communication startup time and per item transfer time, respectively [1].
6. Combining two vectors of length n items requires time $n\gamma$.

3 Combine-to-All

In this section, we discuss strategies for implementing global combine operations that leave the result on all nodes. In all cases, it is assumed that before initiating the combine, each node $\mathbf{P}_i$ owns a vector x_i of length n. Upon completion of the global combine, this vector is overwritten on all nodes with the result of performing an element-wise combine on corresponding elements of the vectors x_i. For simplicity, we will assume n equals a multiple of p in the subsequent discussion.

3.1 Version 1

The first approach we will discuss embeds a minimum spanning tree in the hypercube network, rooted at node $\mathbf{P}_0$. The vectors are pairwise combined until the result is left at the root, after which the result is broadcasted, again utilizing the minimum spanning tree. This process is illustrated for $d = 2$ in Figure 1.

The total time complexity for this strategy equals $2d(\alpha + n\beta) + dn\gamma$.

[1] When a node cannot simultaneously send and receive from the same neighbor, both α and β can be adjusted appropriately.

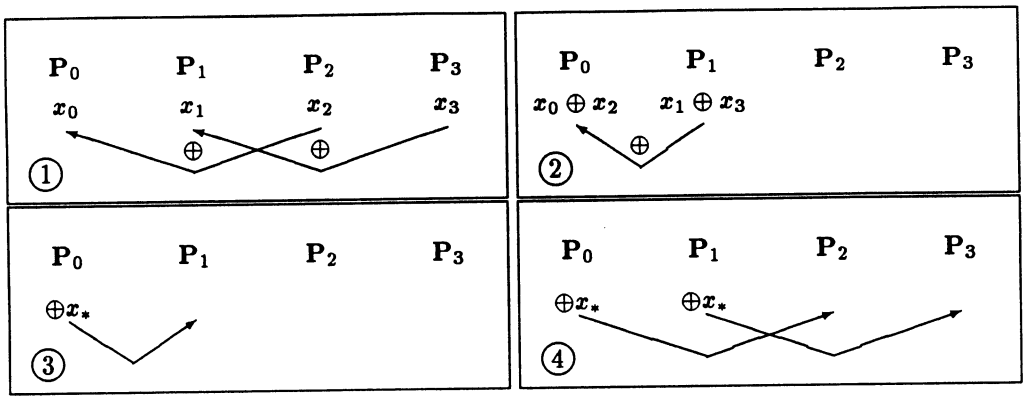

Figure 1: Version 1 on 4 nodes

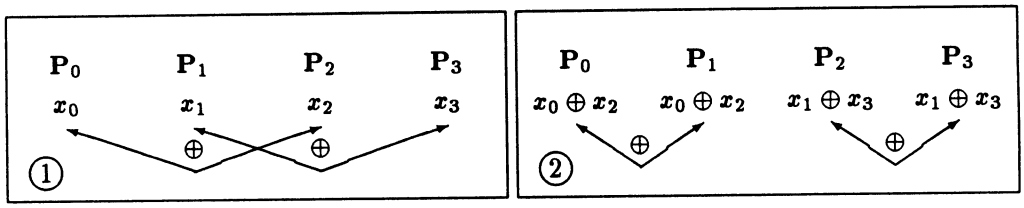

Figure 2: Version 2 on 4 nodes

3.2 Version 2

A second approach, described in [4], Section 14-5.4, starts by each node $\mathbf{P}_i$ exchanging the contents of vector x_i with its neighbor in direction $d-1$, after which the contents of x_i are combined with the contents of the vector that arrived from direction $d-1$, leaving a partial result in x_i. The process proceeds for directions $d-2$ through 0, after which all nodes own a copy of the desired result, as illustrated for $d=2$ in Figure 2.

The total time complexity of this approach equals $d(\alpha + n\beta + n\gamma)$ Notice that under our assumptions, the time complexity of Version 1 always exceeds that of Version 2 by $d(\alpha + n\beta)$.

3.3 Version 3

Neither Version 1 nor Version 2 are optimal in the following sense: even if $\alpha = \beta = 0$, combining p vectors of length n requires time $dn\gamma$ on p nodes versus $(p-1)n\gamma$ on a single node, yielding a speedup of only $(p-1)/\log_2(p)$.

The following approach, discussed for the global vector sum in [4], Chapter 19, does yield optimal use of the nodes if $\alpha = \beta = 0$: Assume n equals an integer multiple of $p = 2^d$. First, each node $\mathbf{P}_i$ divides its vector x_i into two equal subvectors and sends the first half to $\mathbf{P}_j$, where $j = \text{ngbr}(i, d-1)$, if the $(d-1)$st binary digit of i equals 1. Otherwise, the second half is sent to $\mathbf{P}_j$. Next, each node combines the half of vector x_j it receives with the half of vector x_i that was

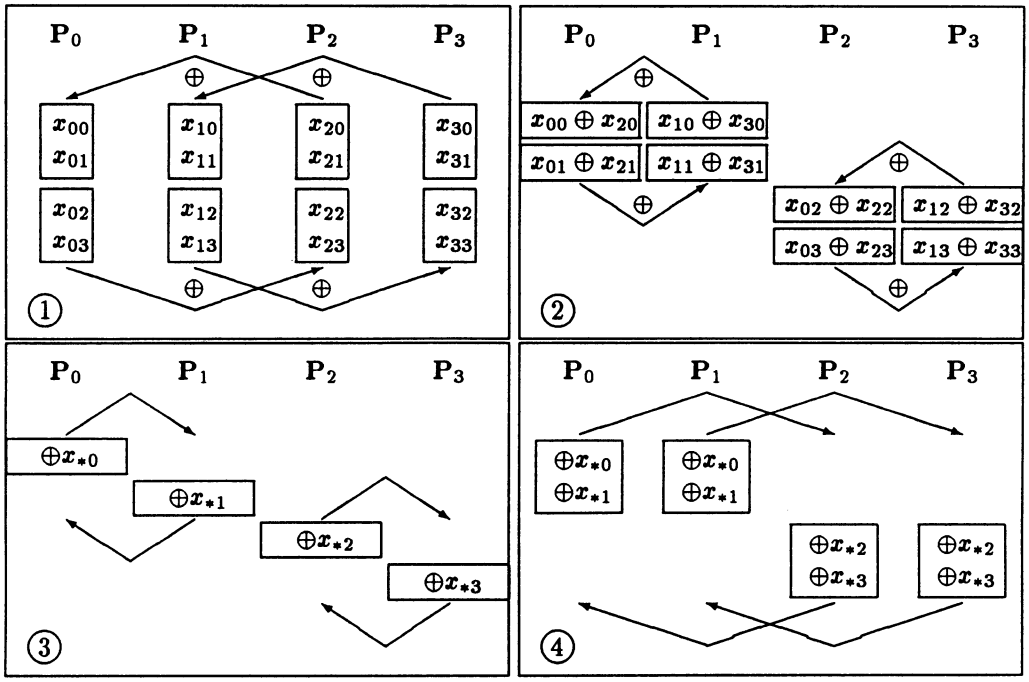

Figure 3: Version 3 on 4 nodes

not sent. This process proceeds for direction $d-2,\ldots,0$, where the size of the partial result being communicated and combined is halved at each step. Finally, each node owns an equal part of the vector that results from combining all vectors x_i. Next, these results are distributed to all nodes by first sending the local results in direction 0, concatenating arriving results with the local results, and proceeding similarly with directions $1,\ldots,d-1$. The whole process is illustrated for $d=2$ in Figure 3. In this figure, each vector x_i is divided into p equal parts, $x_{i0},\ldots,x_{i(p-1)}$.

The total time complexity for Version 3 equals

$$\sum_{i=1}^{d}\left(2\alpha + 2^{-i}n(2\beta+\gamma)\right) = 2d\alpha + \frac{p-1}{p}n(2\beta+\gamma)$$

Depending on the relative values of α, β, and γ, as well as d and n, either Version 2 or Version 3 can yield a faster execution time. This is illustrated in Figure 7 for a cube of dimension 6 with $\alpha = 525\mu$sec, $\beta = 2\mu$sec, and $\gamma = .35\mu$sec, which we shall see are typical parameters when performing combine operations on the Intel iPSC/860.

3.4 Version 4

Notice that in all of the above strategies the global combine can be factored into a combine of vectors on neighboring nodes along direction $(d-1)$, followed by recursively performing global

combines of the partial result on subcubes of dimension $d-1$. This suggests a family of hybrid strategies that combine Versions 2 and 3.

We can generate 2^d separate strategies by considering $S \in \{0,1\}^d = (S_0, \ldots, S_{d-1})$. Here $S_j = 0$ indicates that for the step of the global combine that combines in direction j, node $\mathbf{P}_i$ and $\mathbf{P}_{\text{ngbr}(i,j)}$ exchange current partial results and both combine them as in Version 2, while $S_j = 1$ implies that only half of the current vector is to be exchanged and combined, like in Version 3. Notice that $S = (0, \ldots, 0)$ yields Version 2, while $S = (1, \ldots, 1)$ yields Version 3.

If we assume x_i is initially stored in vector `x = x[0,...,n-1]` on node i, pseudo code that drives $\mathbf{P}_i$ is given in Figure 4.

For a given strategy S, the time complexity of the hybrid strategy is given by the recursion

$$T(S,n,d) = \begin{cases} 0 & \text{if } d < 0 \\ T(S^{d-1}, n, d-1) + \alpha + n\beta + n\gamma & \text{if } S_{d-1} = 0 \\ T(S^{d-1}, \frac{n}{2}, d-1) + 2\alpha + n\beta + \frac{n}{2}\gamma & \text{if } S_{d-1} = 1 \end{cases}$$

where $S^{d-1} = (S_0, \ldots, S_{d-2})$.

4 Optimal Choice of Hybrid Strategy

Naturally, we would like to know how to choose S given vectors of length n and cube dimension d.

Theorem 1 *Given n equal to an integer multiple of p, let k equal the smallest integer so that*

$$n \geq 2^{d-k} \frac{\alpha}{k(\beta+\gamma)+\gamma}$$

Then the optimal choice of hybrid strategy equals $\bar{S}$, where

$$\bar{S}_i = \begin{cases} 0 & \text{if } i < k \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Intuitively, this theorem implies that at first it is benefitial to half the vector length at each stage until the benefit no longer outweighs the cost of the extra communication startup overhead.

The following lemma will aid in understanding the proof of Theorem 1.

Lemma 2 *For any strategy $S \in \{0,1\}^d$ and $m < n$*

$$\begin{aligned} T((0,\ldots,0),n,d) &= d(\alpha + n\beta + n\gamma) \\ T((1,\ldots,1),n,d) &= 2d\alpha + (p-1)n/p(2\beta+\gamma) \\ T(S,m,d) &< T(S,n,d) \\ T((S_0,\ldots,S_{d-1},0,\ldots,0),n,d+d') &= T((0,\ldots,0),n,d') + T(S,n,d) \\ T((1,\ldots,1,S_0,\ldots,S_{d-1}),n,d+d') &= T(S,n,d) + T((1,\ldots,1),\bar{n},d') \end{aligned}$$

for some $\bar{n} \leq n$.

Proof: The first four results follow immediately from the definition of $T(S,n,d)$. The last follows from the observation that the time of the first j steps of hybrid algorithm is independent of the time for the last $d-j$ steps. However, the time is influenced by the length $\bar{n}$ of the partial results when the last $d-j$ steps start. Clearly $\bar{n} \leq n$. □

Proof of Theorem 1: Given n and d, let k and $\bar{S}$ be as defined in Theorem 1. Let $S \in \{0,1\}^d$ be an optimal strategy. We will show that $T(S,n,d) < T(\bar{S},n,d)$ leads to a contradiction.
Case 1: Assume $(d-1) < k$. Then $n < 2\alpha/((d-1)(\beta+\gamma)+\gamma)$ and

$$\begin{aligned} \bar{S} &= (0,0,\ldots,0) \\ S &= (0,\ldots,0,1,S_{j+1},\ldots,S_{d-1}) \end{aligned}$$

for some $j \in \{0,\ldots,d-1\}$. Define strategy R by

$$R = (0,\ldots,0,0,S_{j+1},\ldots,S_{d-1})$$

By Lemma 2,

$$T(R,n,d-1) = T(S,n,d-1) - T(S,m,j+1) + T(R,m,j+1)$$

for some $m \leq n$. The optimality of S implies that

$$T(R,m,j+1) \geq T(S,m,j+1) \tag{2}$$

But since

$$\begin{aligned} T(S,m,j+1) &= 2\alpha + m\beta + \frac{m}{2}\gamma + j(\alpha + \frac{m}{2}(\beta+\gamma)) \\ T(R,m,j+1) &= (j+1)(\alpha + m(\beta+\gamma)) \end{aligned}$$

(2) implies $2\alpha/(j(\beta+\gamma)+\gamma) \leq m$. However, this leads to the contradiction

$$m \leq n < \frac{2\alpha}{(d-1)(\beta+\gamma)+\gamma} \leq \frac{2\alpha}{j(\beta+\gamma)+\gamma} \leq m$$

Case 2: Assume $k \leq d-1$. Let $j \in \{0,\ldots,d-1\}$ equal the largest integer so that $S_j \neq \bar{S}_j$.
Case 2a: $j > k$. Then $S = (S_0,\ldots,S_{j-1},0,1,\ldots,1)$. Define strategy

$$R = (S_0,\ldots,S_{j-1},1,1,\ldots,1)$$

Now

$$\begin{aligned} T(S,n,d) &= \sum_{i=1}^{d-j-1} (2\alpha + 2^{-i}n(2\beta+\gamma)) + \alpha + 2^{-(d-j-1)}(\beta+\gamma) + T(S,2^{-(d-j-1)}n,j) \\ T(R,n,d) &= \sum_{i=1}^{d-j} (2\alpha + 2^{-i}n(2\beta+\gamma)) + T(S,2^{-(d-j)}n,j) \end{aligned}$$

Since $T(S, 2^{-(d-j-1)}n, j) > T(S, 2^{-(d-j)}n, j) \geq 0$ and $T(S, n, d) \leq T(R, n, d)$, we conclude that

$$\alpha + 2^{-(d-j-1)}n(\beta+\gamma) < 2\alpha + 2^{-(d-j)}n(2\beta+\gamma)$$

and hence

$$n < 2^{d-j}\frac{\alpha}{\gamma} < 2^{d-k}\frac{\alpha}{\gamma} \leq 2^{d-k}\frac{\alpha}{k(\beta+\gamma)+\gamma}$$

which contradicts the definition of k.
Case 2b: $j \leq k$. Then $S = (S_0, \ldots, S_k, 1, \ldots, 1)$ and

$$T(S, n, d) = \sum_{i=1}^{d-k-1} (2\alpha + 2^{-k}n(2\beta+\gamma)) + T(S^{k+1}, 2^{-(d-k-1)}n, k+1)$$

However,

$$T(\bar{S}, n, d) = \sum_{i=1}^{d-k-1} \Big(2\alpha + 2^{-k}n(2\beta+\gamma)\Big) + T((0, \ldots, 0), 2^{-(d-k-1)}n, k+1)$$

and, by Case 1 above,

$$T((0, \ldots, 0), 2^{-(d-k-1)}n, k+1) \leq T(S, 2^{-(d-k-1)}n, k+1)$$

which again leads to a contradiction. □

Note 3 *The optimal hybrid algorithm is now derived from the algorithm given in Figure 4 by eliminating* S *from the calling sequence, and replacing the condition in the first* `if` *clause by*

$$\mathtt{n} < 2\alpha/(\mathtt{d}-1)(\beta+\gamma)+\gamma)$$

The time complexity of this optimal hybrid strategy is given by

Corollary 4 *Let n, d, and k be as given in Theorem 1. Then the time complexity of the optimal hybrid strategy equals*

$$2(d-k)\alpha + (1 - 2^{-(d-k)}n(2\beta+\gamma) + k(\alpha + 2^{-(d-k)}n(\beta+\gamma))$$

If $k > d$, k must be replaced by d in this formula.

```
hybridCOMB(n, x, y, d, S)
begin
     if S_{d-1}= 0 [a] then
          send (n, x, ngbr(i,d-1))
          recv (n, y, ngbr(i,d-1))
          combine (n, x, y)
          if d-1 > 0 call hybridCOMB(n, x, y, d-1, S)
     else
          let x0 = x[0,...,n/2-1], x1 = x[n/2,...,n]
          if bit(i,d-1)=0 then
               send(n/2, x1, ngbr(i,d-1))
               recv(n/2, y, ngbr(i,d-1))
               combine(n/2, x0, y)
               if d-1>0 then call hybridCOMB(n/2, x0, y, d-1, S)
               send(n/2, x0, ngbr(i,d-1))
               recv(n/2, x1, ngbr(i,d-1))
          else
               send(n/2, x0, ngbr(i,d-1))
               recv(n/2, y, ngbr(i,d-1))
               combine(n/2, x1, y)
               if d-1>0 call hybridCOMB(n/2, x1, y, d-1, S)
               send(n/2, x1, ngbr(i,d-1))
               recv(n/2, x0, ngbr(i,d-1))
end
```

[a] In Section 4 it will be shown that an optimal hybrid strategy can be obtained by deleting S from the calling sequence and replacing this condition by

$$n < 2\alpha/((d-1)(\beta+\gamma)+\gamma)$$

Figure 4: Hybrid global combine routine

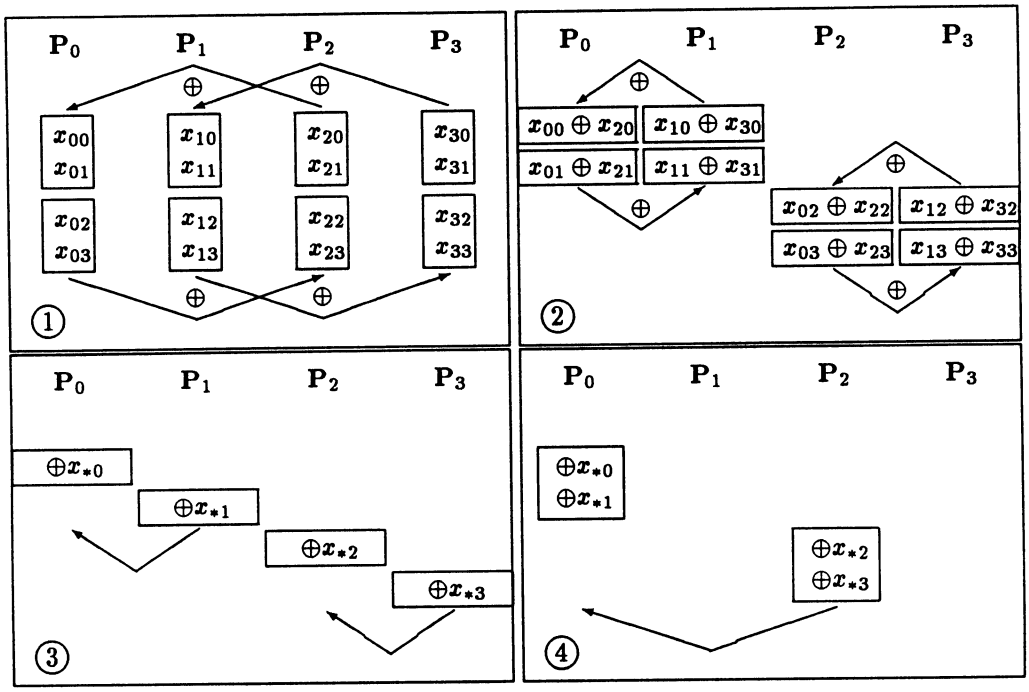

Figure 5: Second approach to combine-to-root on 4 nodes

5 Combine-to-Root

Some algorithms require the result of the global combine to only be known to one node. Without loss of generality, we can take this node to be $\mathbf{P_0}$. Again, there are two commonly used algorithms.

The first approach progresses exactly like Version 1 in the previous section, except that no broadcast of the result to all nodes is necessary (see Steps 1 and 2 of Figure 1). The time complexity for this approach is identical to the time complexity of Version 2 in the previous section.

A second approach is identical to Version 3 in the previous section, except that once each node has computed its part of the global combine, the results are gathered to the root using a minimum spanning tree, as illustrated in Figure 5. The time complexity for this second approach is identical to the time complexity for Version 3.

Naturally, these two approaches can be combined to form a hybrid global combine operation, very much like Version 4 in Section 3. Since the time complexity of the two approaches being combined are identical to those being combined in Version 4, Theorem 1 holds in this case as well, and the optimal hybrid algorithm is given in Figure 6.

```
hybridCOMB2R(n, x, y, d)
begin
     if n < 2α / ((d-1)(β+γ)+γ) then
          if bit(i,d-1)=1 then
               send (n, x, ngbr(i,d-1))
          else
               recv (n, y, ngbr(i,d-1))
               combine (n, x, y)
               if d-1 > 0 call hybridCOMB2R(n, x, y, d-1)
     else
          let x0 = x[0,...,n/2-1], x1 = x[n/2,...,n]
          if bit(i,d-1)=0 then
               send(n/2, x1, ngbr(i,d-1))
               recv(n/2, y, ngbr(i,d-1))
               combine(n/2, x0, y)
               if d-1>0 call hybridCOMB2R(n/2, x0, y, d-1)
               recv(n/2, x1, ngbr(i,d-1))
          else
               send(n/2, x0, ngbr(i,d-1))
               recv(n/2, y, ngbr(i,d-1))
               combine(n/2, x1, y)
               if d-1>0 call hybridCOMB2R(n/2, x1, y, d-1)
               send(n/2, x1, ngbr(i,d-1))
end
```

Figure 6: Optimal hybrid global combine-to-root routine

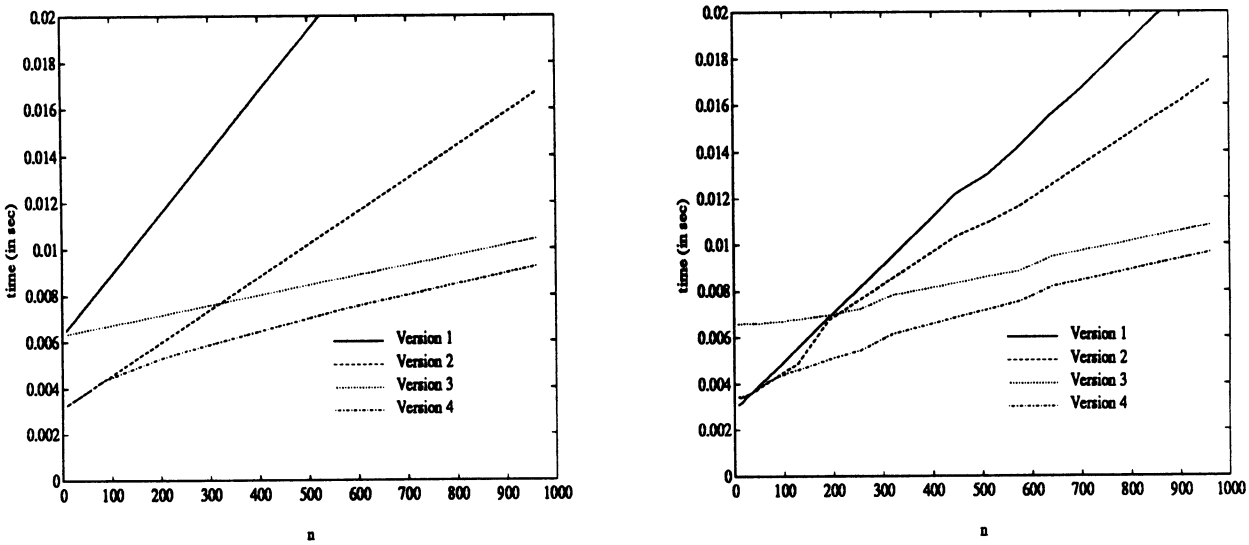

Figure 7: Predicted (*left*) and observed (*right*) time as a function of vector length n on 64 nodes when $\alpha = 525\mu$sec, $\beta = 2\mu$sec, and $\gamma = .35\mu$sec.

6 Experiments on the Intel iPSC/860

To test our theoretical results, we implemented the various global combine operations on the Intel iPSC/860. Our experiments centered around a specific global combine operation, the global summation of single precision floating point vectors.

The Intel iPSC/860 is a commercial parallel processor that consists of Intel i860 processors connected in a hypercube topology. Although this machine is somewhat more general than assumed in Section 2, it can be programmed in such a way that all assumptions are more or less satisfied.

In [3] it is shown that the cost for communicating a floating point vector of length n on the iPSC/860 is roughly given by Assumption 5, with $\alpha = 136\mu$sec and $\beta = 1.6\mu$sec. The cost of adding to single precision floating point numbers is $\gamma = .35\mu$sec. Unfortunately, when vectors of length 25 or less are communicated, the communication startup time becomes $\alpha = 75\mu$sec. To overcome this complication, we padded all communication for our first experiments with an additional 30 floating point numbers. There is some overhead associated with making the subroutine calls and general bookkeeping, yielding an effective $\alpha = 525\mu$sec. We also observed a slightly higher per item overhead: $\beta = 2.0\mu$sec.

The first series of experiments measured the the time required for executing the various approaches described in Section 3. The results are reported in Figure 7. Version 1 has a lower communication startup time than predicted in Figure 7. We believe that this can be attributed to the fact that this algorithm allows some communication and computation to overlap.

Figure 8 shows the performance of the global combine algorithms when the vectors are not padded to compensate for the lower communication startup time when the length of vectors being

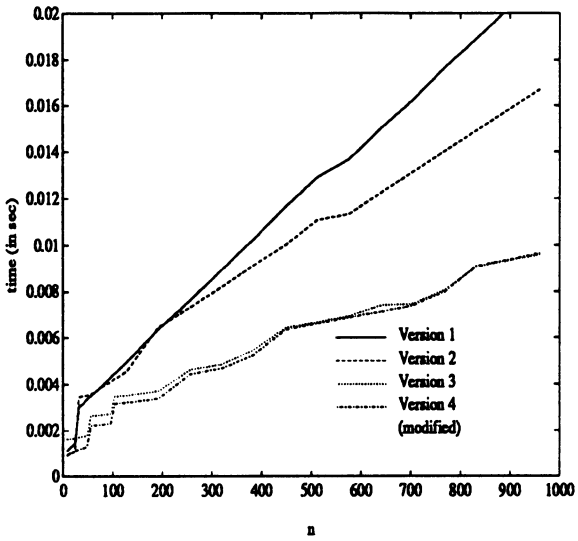

Figure 8: Observed time as a function of vector length n on 64 nodes. In this version, the length of vectors being communicated has not been adjusted, and the algorithm has been modified as described in Section 6, with $\alpha = 480\mu$sec, $\beta = 2\mu$sec, and $\gamma = .35\mu$sec.

communicated is less than 25. While there is a dramatic drop in execution time for Version 1 and 2 when $n \leq 25$, Version 3 is particularly aided by the reduction in startup time, since the vectors being communicated are halved at each step. For example, on a cube of dimension 6, if the original vector is of length 512, the vectors communicated during the last two stages are of length 16 and 8, respectively, thereby reducing the execution time. This drastically affects the optimal choice $\bar{S}$ in Section 4. The following modification appears to give statisfactary results: If $n > 25 \times 2^d$, the optimal choice as described in Section 4 is used. Otherwise, a strategy that halves the vector at each stage until the vector length is less than or equal to 25, and communicates and combines full vectors thereafter, is used. The results are given in Figure 8. Notice that in this case the hybrid method shows significant improvement only for short vectors.

7 Conclusions

The implementation of global combine operations is greatly influenced by the machine on which the operations are to be performed. In this paper, we have proposed a hybrid approach that outperforms standard implementations under the restrictions mentioned in Section 2.

The Intel iPSC/860 failed to meet Assumption 4, requiring a modification to the hybrid algorithm. Other parallel architectures may be more flexible than the assumptions in Section 2. For example, if a hypercube architecture has the feature that it can communicate to all neighbors simultaneously, the hybrid algorithm can be generalized in the following way: At each stage, the vectors

to be combined and communicated are partitioned into d (the dimension of the cube) equal parts, where one part is exchanged in each direction. This kind of approach to increasing the utilization of the communication network is discussed in [4], Chapter 21. The effect is to reduce the constant multiplying the β term in the time complexities of Versions 2 and 3 by a factor d, which can be carried through to reduce the execution time of the hybrid algorithm as well. Details go beyond the scope of this paper.

Acknowledgements

The author would like to thank Dr. Israel Nelken and Dr. Bernard Tourancheau for helpful comments.

References

[1] E. Anderson, A. Benzoni, J. Dongarra, S. Moulton, S. Ostrouchov, B. Tourancheau, and R. van de Geijn. LAPACK for distributed memory architectures: Progress report. In *Proceedings Fifth SIAM Conference on Parallel Processing for Scientific Computing*, Philadelphia, 1991. SIAM.

[2] J.J. Dongarra and R.A. van de Geijn. Reduction to condensed form for the eigenvalue problem on distributed memory architectures. LAPACK Working Note 30, technical report, University of Tennessee, 1991.

[3] T.H. Dunigan. Performance of the Intel iPSC/860 hypercube. Technical Report ORNL/TM-11491, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1990.

[4] G. Fox, M. Johnson, G. Lyzenga, S. Otto, J. Salmon, and D. Walker. *Solving Problems on Concurrent Processors*, volume I. Prentice Hall, 1988.

[5] C. Moler, J. Little, and S. Bangert. *Pro-Matlab, User's Guide.* The Mathworks, Inc., 1987.