Optimal Sorting Algorithms for Parallel Computers

Gerard Baudet

David Stevenson

May, 1975

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania

ABSTRACT:

The problem of sorting a sequence of n elements on a parallel computer with k processors is considered. The algorithms we present can all be run on a single instruction stream multiple data stream computer. For large n, each achieves an asymptotic speed-up ratio of k with respect to the best sequential algorithm. This linear (in k) speed-up is optimal in the number of processors used.

This work was supported by the National Science Foundation under Grant GJ-32111 and the Office of Naval Research under Contract N00014-67-A-0314-0010, NR 044-422.

INTRODUCTION:

From the advent of parallel computers, much attention has been paid to the design of efficient algorithms for these machines. In this paper, we are interested in sorting algorithms for those parallel computers referred to as SIMD machines (single instruction stream multiple data stream machines [Flynn 1972]). In these parallel computers all processors execute the same instruction at the same time, but an instruction may be processing different data for different processors. Illiac IV is such a parallel computer (Barnes et al. [1968]).

We assume (1) that the machine is composed of k processors (P1,P2,...,Pk) and that each processor Pi has its own memory Mi which can be accessed directly by Pi; (2) that a special instruction, the route instruction, allows the processor Pi to read one cell of the memory of an adjacent processor (where adjacent depends upon the interconnection strategy) and store its contents in its own memory, Mi; and (3) that each processor has the capability of inhibiting the execution of the current instruction by setting an appropriate indicator.. These assumptions are not restrictive. In particular, Illiac IV has these properties.

The central idea of this paper is the following: given any algorithm using only comparison-exchanges for sorting k elements with k processors, there is a corresponding algorithm for sorting rk elements with k processors where every comparison in the first algorithm is replaced by a merge-sorting of two ordered lists of r elements in the second. Since the time for merging two ordered lists is also a lower bound for routing the information involved (i.e. both are linear), it follows that the time complexity of routing will determine the complexity of the sorting algorithms in this family, provided the initialization time for sorting the lists of r elements can be neglected.

Because of the effect of routing, we consider three methods for connecting the processors: a linear order; a two-dimensional array scheme (as is used in the Illiac IV); and a "perfect shuffle." For each interconnection pattern there is an rk-element/k-processor sorting algorithm that is optimal in the sense that for suitable choices of r, the speed-up of the parallel sorting over the optimal sequential sorting algorithm is the number of processors used, which, of course, is the maximal possible increase using parallelism. When the maximum speed-ups for the three interconnection methods are expressed in terms of the number of elements sorted, say n, they are, respectively: log n; (log n/log log n)**2; and $2**\sqrt{\log n}$. This is also the number of processors used.

Previous studies in parallel sorting (c.f. Stone [1973]) have frequently equated the number of elements to be sorted with the number of processors. These studies concluded that the speed-up was logarithmic in the number of processors, or, using a special interconnection scheme, n/log n. No algorithm that had a speed-up that was linear in the number of processors was known. While the equation of the number of processors with the number of elements may be reasonable for sorting networks (i.e. an ensemble of logical circuits to sort n inputs), it is likely that users of parallel computers will want to employ more than one memory location per processor. Indeed, one interpretation of the results we derive is that they are an indication of the optimal ratio of memory size to the number of processors (with respect to the interconnection scheme used).

The family of sorting algorithms we present is in itselfs very simple but, so far, no optimal sorting algorithm was known for parallel computers. The best known result was an implementation of Quicksort for sorting a sequence of n elements on a parallel computer with k=log n processors, leading to an average complexity of O(n), as in our first algorithm, but with a worst case complexity of O(n**2/log n) (Stone [1975]). In contrast, our family of parallel sorting algorithms provides optimal speed-up for problems which are large in relation to the number of available processors (where large depends upon the particular interconnection strategy used for processor communication).

1. A NEIGHBORHOOD SORT FOR A LINEARLY CONNECTED SYSTEM:

We begin this section by giving some notation and definitions that are used throughout the paper. An r-sequence, S, is an ordered set of n elements with n=rk, such that the first r elements, S1, are stored in M1, the next r elements, S2, are stored in M2, etc. It will be denoted by S=S1;S2;...;Sk. An r-sequence S=S1;S2;...;Sk is partially sorted if each Si is a sorted sequence. An r-sequence S=S1;S2;...;Sk is sorted if the whole sequence S1,S2,...,Sk is sorted.

The reader should note that this definition of a sorted sequence coincides with the standard notion of a sorted sequence (i.e. as memory address increases the sequence is monotone) only when each memory Mi contains consecutive memory locations. Another way of addressing memory is to have in memory Mi all addresses which modulo k are congruent to i (as is done for example in the Illiac IV). For this case, a sorted r-sequence will not be sorted with respect to memory location, but a post-processing step which accomplishes this re-ordering can be performed in time linear in the number of elements sorted. Since this does not affect the nature of our asymptotic results, we shall not treat the matter further.

In this section we will present and analyze a sorting algorithm for a machine for which the processors adjacent to Pi are Pj where j=i-1 mod k or j=i+1 mod k. The algorithm takes, as input, an arbitrary r-sequence, and gives as output the corresponding sorted r-sequence. It is illustrated in figure 1.

P1 / M1 :	43	63	54	P1 / M1 :	17	25	28
P2 / M2 :	28	79	72	P2 / M2 :	32	43	47
P3 / M3 :	32	47	84	P3 / M3 :	54	63	66
P4 / M4 :	66	25	17	P4 / M4 :	72	79	84
(a) Input				(b) Output			

Figure 1: Effects of the algorithm on the 3-sequence
43,63,54;28,79,72;32,47,84;66,25,17.

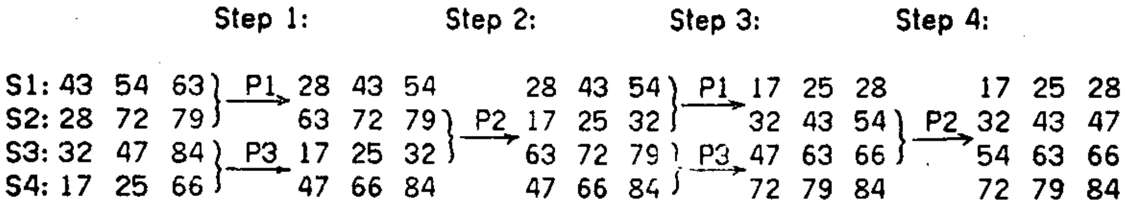
The algorithm is based upon a generalization of the neighbor sort, (Habermann

[1972]), and will be described in section 1.1. In section 1.2. we will give an implementation of this algorithm for the machine we consider. An analysis of the algorithm will be presented in section 1.3. Taking into account both the computational cost, measured as the number of comparisons executed, and the routing cost, the cost for transferring the data between different memories, it will be shown that the algorithm takes time (n log n)/k + O(n). Therefore, when k < log n the asymptotic speed-up ratio is k and this is optimal in the number of processors used.

1.1. THE ALGORITHM:

Habermann has presented an algorithm for sorting k elements using k processors in k steps of parallel "comparison exchanges." The algorithm is the following. Let a1, a2, ..., ak be the sequence to be sorted. In the first step, for i=1,3,...,2⌊k/2⌋-1, processor Pi compares elements ai and ai+1 and if ai > ai+1 the two elements are exchanged. In the second step, the same comparison exchanges are executed for i=2,4,...,2⌊(k-1)/2⌋. Steps 3, 5, ... are the repetitions of step 1. And steps 4, 6, ... are the repetitions of step 2. The proof that this algorithm leads to a completely sorted sequence after at most k steps is based on the observation that the distance between the position of an element in the sequence after p steps and its final position is bounded by k-p.

Now we give a generalization of this algorithm to partially sorted r-sequences. In the first step, for i=1,3,...,2⌊k/2⌋-1, processor Pi merges the two sub-sequences Si and Si+1 and then assigns to Si the first half of the resulting merged sequence (i.e., the r smallest elements) and assigns to Si+1 the second half (i.e., the r largest elements). For the second step, the same operations are executed but for i=2,4,...,2⌊(k-1)/2⌋. Again steps 3, 5, ... are repetitions of step 1, and steps 4, 6, ... are repetitions of step 2. This is illustrated in figure 2.



	Step 1:		Step 2:		Step 3:		Step 4:	
S1: 43 54 63	P1	28 43 54		28 43 54	P1	17 25 28		17 25 28
S2: 28 72 79		63 72 79	P2	17 25 32		32 43 54	P2	32 43 47
S3: 32 47 84	P3	17 25 32		63 72 79	P3	47 63 66		54 63 66
S4: 17 25 66		47 66 84		47 66 84		72 79 84		72 79 84

Figure 2: The four steps of parallel "merging-splittings"
for sorting the partially sorted 3-sequence
43,54,63;28,72,79;32,47,84;17,25,66.

A proof can be given that the algorithm leads to a sorted r-sequence in at most k steps. It is based on the same observation and can be adapted from the proof given by Habermann.

1.2. IMPLEMENTATION OF THE ALGORITHM:

Assume we know how to merge two sorted sequences Ai and Bi of equal length using processor Pi, and that it can be done in parallel for all processors (that is, the control structure of the algorithm must depend upon data-dependent masks rather than upon data-dependent branches). In particular, the merge sorting algorithm (see for example Knuth [1973]) is an ideal method for completing the first phase, that is, for initially sorting the Si. The second phase, sorting the partially sorted r-sequence using the k steps of parallel "merge splitting" as described above, can also be done easily on an SIMD machine, as can be seen in the program given below.

The program is written in an Algol-like language using an additional feature for describing the parallelism. Namely, when a statement is followed by a mask, only those processors Pi for which bit i of the mask is set to 1 will execute the instruction, the other processors being inhibited. When no specification follows a statement every processor is to execute the corresponding instruction.

The data structure involved in the program includes in memory Mi, for i=1,2,...k, an array Si[1:r] containing the current sub-sequence Si of the r-sequence to be sorted, and three arrays Ai[1:r+1], Bi[1:r+1] and Ci[1:2r] used as working areas. For simplicity we assume that k, the number of processors, is even. We use two masks: in EVEN the bit i is set to 1 only for i=2,4,...,k-2, and in ODD only for i=1,3,...,k-1. These masks are used only for the implementation of the second phase.

The procedure MERGE, when executed by processor Pi with parameter p, merges the two subsequences Ai[1:p] and Bi[1:p] (that is the first p elements in both arrays Ai and Bi) and stores the resulting merged sequence in Ci[1:2p].

```
MERGE(p)=
begin
Ai[p+1]←w;Bi[p+1]←w;
a←1;b←1;
for c←1 step 1 until 2p do
   if Ai[a]>Bi[b]
   then begin
      Ci[c]←Bi[b];b←b+1
      end
   else begin
      Ci[c]←Ai[a];a←a+1
      end;
end
```

comment: w is added as a sentinel at the end of both sequences Ai and Bi so that no supplementary test is required for checking if any sequence becomes exhausted. (w is supposed to be greater than any element that Ai and Bi can ever contain.)

comment: We use implicitly two masks which are set according to the result of the comparison between A[a] and B[b]. (This is easily done if there exists an instruction which allows a processor to set its inhibition indicator according to its condition code.)

```
Phase 1:
for p←1 step p until r-1 do
   for j←1 step 2p until r do
      begin
      Ai[1:p]←Si[j:j+p-1];
      Bi[1:p]←Si[j+p:j+2p-1];
      MERGE(p);
      Si[j:j+2p-1]←Ci[1:2p];
      end;
```

comment: The phase 1 as it is programmed here is able to partially sort only those r-sequences for which r is a power of 2. This restriction can be easily removed if we add artificial elements.

comment: The phase 1 is completed at this point and S1[1:r], S2[1:r], ... Sk[1:r] contain the partially sorted r-sequence.

```
Phase 2:
for j←1 step 2 until k-1 do
   begin
   Ai[1:r]←Si[1:r];          mask EVEN;
   Bi[1:r]←Si+1[1:r];        mask EVEN;
   MERGE(r);                 mask EVEN;
   Si[1:r]←Ci[1:r];          mask EVEN;
   Si+1[1:r]←Ci[r+1:2r];     mask EVEN;
   Ai[1:r]←Si[1:r];          mask ODD;
   Bi[1:r]←Si+1[1:r];        mask ODD;
   MERGE(r);                 mask ODD;
   Si[1:r]←Ci[1:r];          mask ODD;
   Si+1[1:r]←Ci[r+1:2r];     mask ODD;
   end
```

comment: This statement e.g. implicitely uses the route instruction.

comment: After those k steps of parallel "merging-splittings", S1[1:r], S2[1:r], ... Sk[1:r] contain the sorted r-sequence, and our goal is achieved.

1.3. ANALYSIS OF THE ALGORITHM:

We first evaluate the routing time. Routing is introduced only by instructions of the form $Bi[1{:}r] \leftarrow Si+1[1{:}r]$ and $Si+1[1{:}r] \leftarrow Ci[r+1{:}2r]$ in the second phase. Such instructions are executed k times each and, in both cases, r route instructions are required, therefore the total number of route instructions is

$$R = 2rk = 2n.$$

For the computational cost, we count only the number of comparison instructions executed but, clearly, the total cost excluding routing has the same order of magnitude. Comparison instructions occur only in the MERGE procedure and 2p comparisons are required when the procedure is called with parameter p. Looking at phase 1, the first for-loop is executed log r times (when r is a power of 2) for $p = 1, 2, 4, \ldots r/2$ and the inner loop is executed $r/2p$ times for each of the values of p. Each of the inner loops requires 2p comparisons, therefore the total number of those instructions is

$$C1 = [2p(r/2p)]\log r = r \log r = (n/k) \log(n/k).$$

For the second phase, the number of comparisons is simply

$$C2 = 2rk = 2n.$$

Therefore the total number of comparisons is

$$C = C1 + C2 = (n \log n)/k - (n \log k)/k + 2n.$$

Assume, now, that the time for executing a comparison instruction is a unit of time. Let λ be the time for executing a route instruction involving an adjacent processor. Then, when we consider only comparisons and routing, the total execution time of the algorithm is

$$T = C + \lambda R = (n \log n)/k - (n \log k)/k + 2n + 2\lambda n$$

or simply

$$T = (n \log n)/k + O(n).$$

But we know that the minimum number of comparisons required for sorting a sequence of n elements on a sequential computer is asymptotically n log n. Therefore, when k is smaller than log n (in order of magnitude), the asymptotic speed-up ratio of our algorithm over the optimal sequential algorithm is k, which is optimal. In particular, when $k=\log n$, the ratio of this parallel algorithm to the optimal sequential algorithm is of order log n, the number of processors.

On the other hand, when k is greater than log n, the total execution time required for our algorithm is linear in n; but this is the best we can achieve with the constraints of the machine we consider here, as can be seen from the following remark. In case all elements whose final destination is memory Mk are initially in the same memory Mi, the total routing must be at least $r(k-i)$ (or $\min\{r(k-i),ri\}$ if we consider processors P1 and Pk adjacent). If we choose, for example, $i=k/2$ then the total routing must be at least n/2 and consequently the total execution time must also be at least linear in n. Therefore, in this case our algorithm is within a constant factor of the optimum time.

2. OTHER OPTIMAL ALGORITHMS:

It seems unlikely that systems with many processors will have the routing time between Pi and Pj linear in |i-j|. Indeed, this is not the case with Illiac IV, which has sixty-four processing elements. In this section we examine two other schemes for connecting processors which are in some sense more efficient in routing information among the processors. The first, a two-dimensional array, is the technique used in Illiac IV; the second, based on the "perfect shuffle," is a design that has been frequently advocated for interconnecting networks of arithmetic processors. We begin with some general observations about our family of parallel sorting algorithms before considering the particular impact of the various interconnection structures.

Our linear-connected parallel algorithm is based on the simple idea of replacing the sequence of parallel "comparison-exchanges" of the initial neighbor sort by a corresponding sequence of parallel "merging-splittings." This idea can be easily generalized. For any algorithm (based exclusively on comparison exchanges) for sorting a sequence of k elements on a parallel computer with k processors, there corresponds another parallel sorting algorithm for sorting a sequence of n=rk elements on a parallel computer with comparable interconnections. In addition, if the initial algorithm requires C(k) parallel comparisons and a total routing R(k), the running time of the new algorithm is, assuming that comparison takes one unit of time and that λ is an average for the time required for a unit step of routing:

$$T(n) = \frac{n \log n}{k} + n\left[\frac{2C(k)-\log k}{k} + \lambda\,\frac{R(k)}{k}\right].$$

Provided that C(k) and R(k) are less than log n (in order of magnitude), the asymptotic speed-up ratio is k which is optimal.

For example, this scheme could be applied to Batcher's sorting algorithm (Batcher [1968]). The resulting algorithm would lead to an improvement over the neighbor algorithm for the number of comparisons (C(k) is reduced from k to (log k)**2) but, for a parallel computer with a linear interconnection scheme, the total routing becomes 2n log k + O(n) as opposed to 2n for the neighbor sort algorithm and thus Batcher's sorting algorithm is not optimal. Hence the computational efficiencies of this algorithm are lost in the routing inefficiencies. This illustrates the often overlooked aspect of parallel computing: that it is the environment in which an algorithm will be executed that determines whether the computational aspects will be realized.

If we take into account other specific interconnection schemes for a parallel computer, it is possible to decrease the time for sorting n elements and hence to increase the speed-up of parallel sorting over sequential sorting. In Illiac IV, the interconnection scheme is such that each processor can access in one route step not only the memory of an adjacent processor but also the memory of a processor at a distance of $\sqrt{k}$. In this case simple (but tedious) algebra shows that the total routing time for Batcher's algorithm is given by

$$R(k) = 4\sqrt{k}.\log k - 8\sqrt{k} - \log k + 8.$$

This shows that asymptotically (for large values of k) an adaptation of

Batcher's sorting algorithm on a two-dimensional array computer will give a better result both for the total number of parallel comparisons (for the second order term only) and for the total routing than will the neighbor sort algorithm on a linearly connected processor. However the following table shows that for small values of k (the number of processors), the total routing of both algorithms are very close. In this table we have reported the quantities R(k)/k for both algorithms, namely RN(k) = (2k)/k=2 in the case of the neighbor sort and RB(k) = $(4\sqrt{k} \log k - 8\sqrt{k} - \log k + 8)/k$ for the adaptation of Batcher's sorting algorithm to exploit the two dimensional connections.

k	=	4	16	64	256	1024
RN(k)	=	2	2	2	2	2
RB(k)	=	1.50	2.25	2.03	1.47	0.99

Table 1: Comparison of the total routing on a square array for the Neighbor sort and Batcher's algorithms.

With this two-dimensional interconnection scheme, the adaptation of Batcher's sorting algorithm achieves an asymptotic speed-up ratio of k as long as C(k) and R(k) are less than log n (in order of magnitude), that is, as long as vk log k < log n. This holds for k less than (log n/log log n)**2. Thus, using this processor configuration, the maximum possible speed-up over the optimal sequential sorting algorithm is greater than log n (the case for the linear interconnection scheme). This improvement in speed-up is due, of course, from using proportionately more processors, and is still optimal in the number of processors used.

And in fact for the actual configuration of Illiac IV, with k=64 and λ approximately 3/4, we find that for r=1024, Batcher's sort is almost twice as fast as the neighbor sort, counting only comparisons and routing.

Finally we briefly consider the effect of using a "perfect shuffle" to connect the processors (Knuth[1973]). In this scheme the processors "adjacent" to Pi are Pj where j=2i mod k and 2j=i mod k. For k a power of two, a sort of k elements can be done with C(k)=R(k)=(log k)**2 (Knuth [1973]). As above, the corresponding merge-splitting algorithm has linear speed-up (in the number of processors) as long as k is less than 2**$\sqrt{\log n}$.

ACKNOWLEDGEMENTS:

We wish to thank J. F. Traub and D. E. Heller, CMU, for comments on the manuscript, and H. T. Kung, CMU, for suggesting to the first author that sorting on a parallel computer would be an interesting problem to study, and for many comments and discussions.

BIBLIOGRAPHY:

Barnes, G. H., et al. [1968]. The ILLIAC IV Computer, IEEE Transactions on Computers, Vol. C-17, no. 8, August 1968, pp. 746-757.

Batcher, K. E. [1968]. Sorting Networks and Their Applications, Spring Joint Computer Conference, AFIPS proceedings, Vol. 32, 1968, pp. 307-314.

Bouknight, W. J., et al., [1972]. The ILLIAC IV System, Proc IEEE, 60, no. 4, April 1972, pp. 369-388.

Flynn, M. J. [1972]. Some Computer Organizations and Their Effectiveness, IEEE Transactions on Computers, Vol. C-21, no. 9, September 1972.

Habermann, A. N. [1972]. Parallel Neighbor Sort, Computer Science Department, Carnegie-Mellon University, August 1972.

Knuth, D. E. [1973]. The Art of Computer Programming, Vol. 3, Sorting and Searching, Addison Wesley, 1973.

Orcutt, S. E. [1974]. Computer Organization and Algorithms for Very High Speed Computations, Ph.D. dissertation, Computer Science Department, Stanford University, September 1974.

Slotnick, D. L., et al., [1962]. The Soloman Computer -- A Preliminary Report, Proceedings of 1962 Workshop on Computer Organization. Washington D.C.: Spartan, 1963, p. 66.

Stone, H. S. [1973]. Problems of Parallel Computations, Proceedings of the Symposium on Complexity of Sequential and Parallel Numerical Algorithms, J. F. Traub ed., New-York, New-York, Academic Press, 1973.

--- [1975]. Private communication.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
OPTIMAL SORTING ALGORITHMS FOR PARALLEL COMPUTERS

5. TYPE OF REPORT & PERIOD COVERED
Interim

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Gerard Baudet and David Stevenson

8. CONTRACT OR GRANT NUMBER(s)
N00014-67-A-0314-0010
Nr 044-422

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Carnegie-Mellon University
Computer Science Dept.
Pittsburgh, PA 15213

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Arlington, VA 22217

12. REPORT DATE
May 1975

13. NUMBER OF PAGES
11

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release- distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The problem of sorting a sequence of n elements on a parallel computer with k processors is considered. The algorithms we present can all be run on a single instruction stream multiple data stream computer. For large n, each achieves an asymptotic speec-up ratio of k with respect to the best sequential algorithm. This linear (in k) speed-up is optimal in the number of processors used.