Scans as Primitive Parallel Operations

Guy E. Blelloch
Carnegie Mellon University

November 1989

Abstract

In most parallel random-access machine (P-RAM) models, memory references are assumed to take unit time. In practice, and in theory, certain scan operations, also known as prefix computations, can executed in no more time than these parallel memory references. This paper outline an extensive study of the effect of including in the P-RAM models, such scan operations as unit-time primitives. The study concludes that the primitives improve the asymptotic running time of many algorithms by an $O(\lg n)$ factor, greatly simplify the description of many algorithms, and are significantly easier to implement than memory references. We therefore argue that the algorithm designer should feel free to use these operations as if they were as cheap as a memory reference.

This paper describes five algorithms that clearly illustrate how the scan primitives can be used in algorithm design: a *radix-sort* algorithm, a *quicksort* algorithm, a *minimum-spanning-tree* algorithm, a *line-drawing* algorithm and a *merging* algorithm. These all run on an EREW P-RAM with the addition of two scan primitives, and are either simpler or more efficient than their pure P-RAM counterparts.

The scan primitives have been implemented in microcode on the Connection Machine System, are available in PARIS (the parallel instruction set of the machine), and are used in a large number of applications. All five algorithms have been tested, and the radix-sort is the currently supported sorting algorithm for the Connection Machine.

Acknowledgement of Financial Support

This report describes research done within the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. Support for the A.I. Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under the Office of Naval Research contract N00014–85–K–0124.

Biography

Guy E. Blelloch received a B.A. degree in physics in 1983 from Swarthmore College, Swarthmore, PA, and an M.S. and a Ph.D. Degree in computer science in 1986 and 1988, respectively, from the Massachusetts Institute of Technology, Cambridge, MA.

He is an Assistant Professor at Carnegie Mellon University, Pittburgh, PA. His research interests include data-parallel algorithms, languages, and compilers. He has worked or consulted for Thinking Machines Corporation since 1985.

Address

Guy Blelloch
Assistant Professor
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
(412) 268-6245

keywords

Scan, Parallel Algorithms, Parallel Computing, Prefix Computations, PRAM, Connection Machine.

1 Introduction

Algorithmic models typically supply a simple abstraction of a computing device and a set of primitive operations assumed to execute in a fixed "unit time". The assumption that primitives operate in unit time allows researchers to greatly simplify the analysis of algorithms, but is never strictly valid on real machines: primitives often execute in time dependent on machine and algorithm parameters. For example, in the serial random access machine (RAM) model [14], memory references are assumed to take unit time even though the data must fan-in on any real hardware and therefore take time that increases with the memory size. In spite of this inaccuracy in the model, the unit-time assumption has served as an excellent basis for the analysis of algorithms.

In the parallel random access machine (P-RAM) models [16, 40, 42, 19, 20], memory references are again assumed to take unit time. In these parallel models, this "unit time" is large since there is no practical hardware known that does better than deterministic $O(\lg^2 n)$, or probabilistic $O(\lg n)$, bit times for an arbitrary memory reference from n processors.[1] This can lead to algorithms that are practical in the model but impractical on real machines. One solution is to use lower-level models based on a fixed connectivity of the processors, such as the shuffle-exchange networks [44] or grid networks [47]. This, however, gives up machine independence and greatly complicates the description of algorithms. This paper suggests another solution: to add other primitives to the P-RAM models that can execute as fast as memory references in practice, and that can reduce the number of program steps of algorithms—therefore making the algorithms more practical.

This paper outlines a study of the effect of including certain scan operations as such "unit time" primitives in the P-RAM models. The *scan operations*[2] take a binary operator $\oplus$ with identity i, and an ordered set $[a_0, a_1, ..., a_{n-1}]$ of n elements, and returns the ordered set $[i, a_0, (a_0 \oplus a_1), ..., (a_0 \oplus a_1 \oplus ... \oplus a_{n-2})]$. In this paper we only consider two primitive scan operators, integer addition and integer maximum. These have a particularly simple implementation and they can be used to implement many other useful scan operations. On a P-RAM, each element a_i is placed in a separate processor, and the scan executes over a fixed order of the processors—the prefix operation on a linked list [48, 27], and the fetch-and-op type instructions [21, 20, 37] are not considered.

[1] The AKS sorting network [1] takes $O(\lg n)$ time deterministically, but is not practical.

[2] The appendix gives a short history of the scan operations.

	Model		
Algorithm	EREW	CRCW	Scan
Graph Algorithms (n vertices, m edges, m processors)			
Minimum Spanning Tree	$O(\lg^2 n)$	$O(\lg n)$	$O(\lg n)$
Connected Components	$O(\lg^2 n)$	$O(\lg n)$	$O(\lg n)$
Maximum Flow	$O(n^2 \lg n)$	$O(n^2 \lg n)$	$O(n^2)$
Maximal Independent Set	$O(\lg^2 n)$	$O(\lg^2 n)$	$O(\lg n)$
Biconnected Components	$O(\lg^2 n)$	$O(\lg n)$	$O(\lg n)$
Sorting and Merging (n keys, n processors)			
Sorting	$O(\lg n)$	$O(\lg n)$	$O(\lg n)$
Merging	$O(\lg n)$	$O(\lg \lg n)$	$O(\lg \lg n)$
Computational Geometry (n points, n processors)			
Convex Hull	$O(\lg n)$	$O(\lg n)$	$O(\lg n)$
Building a K-D Tree	$O(\lg^2 n)$	$O(\lg^2 n)$	$O(\lg n)$
Closest Pair in the Plane	$O(\lg^2 n)$	$O(\lg n \lg \lg n)$	$O(\lg n)$
Line of Sight	$O(\lg n)$	$O(\lg n)$	$O(1)$
Matrix Manipulation ($n \times n$ matrix, n^2 processors)			
Matrix $\times$ Matrix	$O(n)$	$O(n)$	$O(n)$
Vector $\times$ Matrix	$O(\lg n)$	$O(\lg n)$	$O(1)$
Linear Systems Solver (with pivoting)	$O(n \lg n)$	$O(n \lg n)$	$O(n)$

Table 1: The scan model is the EREW P-RAM mode with the addition of two scan primitives. These scan primitives improve the asymptotic running time of many algorithms by an $O(\lg n)$ factor. The algorithms not described in this paper are described elsewhere [7, 8]. Some of the algorithms are probabilistic.

	Memory Reference	Scan Operation
Theoretical		
VLSI models		
Time	$O(\lg n)$ [29]	$O(\lg n)$ [30]
Area	$O(n^2/\lg n)$	$O(n)$
Circuit models		
Depth	$O(\lg n)$ [1]	$O(\lg n)$ [15]
Size	$O(n \lg n)$	$O(n)$
Actual		
64K processor CM-2		
Bit Cycles (Time)	600	550
Percent of Hardware	30	0

Table 2: Both in theory and in practice certain scan operations can execute in less time than references to a shared memory, and can be implemented with less hardware. For the CM-2 implementation, the scan is implemented in microcode and only uses existing hardware.

The conclusions of our study are summarized as follows.

- The scan primitives improve the asymptotic running time of many algorithms by an $O(\lg n)$ factor over the EREW model and some by an $O(\lg n)$ factor over the CRCW model (see Table 1).

- The scan primitives simplify the description of many algorithms. Even in algorithms where the complexity is not changed from the pure P-RAM model, the scan version is typically significantly simpler.

- Both in theory and in practice, the two scan operations can execute in less time than references to a shared memory, and can be implemented with less hardware (see Table 2).

This paper is divided into two parts: *algorithms* and *implementation*. The first part illustrates how the scan primitives can be used in algorithm design, describes several interesting algorithms, and proves some of the results of Table 1. The second part describes a very simple hardware implementation of the two scan primitives, and describes how other scan operations can be implemented on top of these two scans.

We call the exclusive-read exclusive-write (EREW) P-RAM model with the scan operations included as primitives, the *scan model*. Since the term *unit time* is misleading (both memory references and scan operations take many clock cycles on a real machine), this paper henceforth uses the term *program step* or *step* instead. The number of program steps taken by an algorithm is the *step complexity*.

2 Algorithms

On first appearance, the scan operations might not seem to greatly enrich the memory reference operations of the P-RAM models. As we will discover in this paper, this is far from true; they are useful for a very broad set of algorithms. We separate the uses of scans into four categories. Section 2.2, *simple operations*, shows how scans can be used to enumerate a subset of marked processors, to copy values across all processors and to sum values across all processors. As an illustration of the use of enumerating, we describe a practical radix-sort that requires $O(1)$ program steps for each bit of the keys. Section 2.3, *segmented operations*, shows how segmented versions of the scans are useful in algorithms that work over many

Uses of Scan Primitives	Example Algorithms
Enumerating (2.2)	Splitting, Load Balancing
Copying (2.2)	Quicksort, Line Drawing, Minimum Spanning Tree
Distributing Sums (2.2)	Quicksort, Minimum Spanning Tree
Splitting (2.2.1)	Split Radix Sort, Quicksort
Segmented Primitives (2.3)	Quicksort, Line Drawing, Minimum Spanning Tree
Allocating (2.4)	Line Drawing, Halving Merge
Load-Balancing (2.5)	Halving Merge

Example Algorithms	Uses of Scan Primitives
Split Radix Sort (2.2.1)	Splitting
Quicksort (2.3.1)	Splitting, Distributing Sums, Copying, Segmented Primitives
Minimum Spanning Tree (2.3.3)	Distributing Sums, Copying, Segmented Primitives
Line Drawing (2.4.1)	Allocating, Copying, Segmented Primitives
Halving Merge (2.5.1)	Allocating, Load Balancing

Table 3: A cross reference of the various uses of scans introduced in this paper with the example algorithms discussed in this paper. All the uses can be executed in a constant number of program steps.

sets of data in parallel. As examples of the use of segmented scans, we describe a quicksort algorithm, which has an expected complexity of $O(\lg n)$, and a minimum-spanning-tree algorithm with probabilistic complexity $O(\lg n)$. Section 2.4, *allocating*, shows how the scan operations are very useful for allocating processors. As an example, we describe a line-drawing algorithm which uses $O(1)$ program steps. Section 2.5, *load-balancing*, shows how the scan operations are useful to load balance elements across processors when there are more data elements than processors. As an examples, we describe a merging algorithm which with p processors and n elements has a step complexity of $O(n/p + \lg n)$.

Table 3 summarizes the uses of the scan operations and the example algorithms discussed in this paper. All the algorithms discussed in the paper have been implemented on the Connection Machine and in some of descriptions we mention the running times of the implementation. Before discussing the uses of the scan primitives, we introduce some notational conventions used in the paper.

2.1 Notation

We will assume that the data used by the algorithms in this paper is stored in vectors (one dimensional arrays) in the shared memory and that each processor is assigned to one element of the vector. When executing an operation, the i^{th} processor operates on the i^{th} element of a vector. For example, in the operation:

$$\begin{array}{lcllllllll}
\mathrm{A} & = & [5 & 1 & 3 & 4 & 3 & 9 & 2 & 6] \\
\mathrm{B} & = & [2 & 5 & 3 & 8 & 1 & 3 & 6 & 2] \\
\mathrm{C} \leftarrow \mathrm{A} + \mathrm{B} & = & [7 & 6 & 6 & 12 & 4 & 12 & 8 & 8]
\end{array}$$

each processor reads its respective value from the vectors A and B, sums the values, and writes the result into the destination vector C. Initially, we assume that the P-RAM always has as many processors as vector elements.

The *scan* primitives can be used to scan the elements of a vector. For example:

$$\begin{array}{lcllllllll}
\mathrm{A} & = & [2 & 1 & 2 & 3 & 5 & 8 & 13 & 21] \\
\mathrm{C} \leftarrow \texttt{+-scan}(\mathrm{A}) & = & [0 & 2 & 3 & 5 & 8 & 13 & 21 & 34]
\end{array}$$

In this paper we only use five primitive scan operations: `or-scan`, `and-scan`, `max-scan`, `min-scan` and `+-scan`. We also use *backward* versions of each of these scans operations—

versions that scan from the last element to the first. Section 3.4 shows that all the scans can be implemented with just two scans, a `max-scan` and a `+-scan`.

To reorder the elements of a vector, we use the `permute` operation. The `permute` operation, in the form `permute`(A, I), permutes the elements of A to the positions specified by the indices of I. All indices of I must be unique. For example:

$$
\begin{array}{lcl}
\text{A (data vector)} & = & [a_0 \quad a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5 \quad a_6 \quad a_7] \\
\text{I (index vector)} & = & [2 \quad 5 \quad 4 \quad 3 \quad 1 \quad 6 \quad 0 \quad 7] \\
\text{C} \leftarrow \texttt{permute}(\text{A, I}) & = & [a_6 \quad a_4 \quad a_0 \quad a_3 \quad a_2 \quad a_1 \quad a_5 \quad a_7]
\end{array}
$$

To implement the `permute` operation on a EREW P-RAM, each processor reads its respective *value* and *index*, and writes the *value* into the *index* position of the destination vector.

2.2 Simple Operations

We now consider three simple operations that are based on the scan primitives: enumerating, copying and distributing sums (see Figure 1). These operations are used extensively as part of the algorithms we discuss in this paper and all have an step complexity of $O(1)$. The `enumerate` operation returns the integer i to the i^{th} true element. This operation is implemented by converting the flags to 0 or 1 and executing a `+-scan`. The `copy` operation copies the first element over all elements. This operation is implemented by placing the identity element in all but the first element of a vector and executing any scan.[3] Since the scan is not inclusive, we must put the first element back after executing the scan. The `+-distribute` operation returns to each element the sum of all the elements. This operation is implemented using a `+-scan` and a backward `copy`. We can likewise define a `max-distribute`, `min-distribute`, `or-distribute` and `and-distribute`.

[3]One might think of defining a binary associative operator `first` which returns the first of its two arguments, and use it to execute the `copy` operation. The problem is that the `first` operator does not have an identity—a requirement for our definition of a scan.

2.2.1 Example: Split Radix Sort

To illustrate the use of the scans for enumerating, consider a simple radix sorting algorithm. The algorithm is a parallel version of the standard serial radix sort [26].

The algorithm loops over the bits of the keys, starting at the lowest bit, executing a `split` operation on each iteration. The `split` operation packs the keys with a 0 in the corresponding bit to the bottom of a vector, and packs the keys with a 1 in the bit to the top of the same vector. It maintains the order within both groups. The sort works because each `split` operation sorts the keys with respect to the current bit (0 down, 1 up) and maintains the sorted order of all the lower bits since we iterate from the bottom bit up. Figure 2 shows an example of the sort along with code to implement it.

We now consider how the `split` operation can be implemented in the scan model. The basic idea is to determine a new index for each element and permute the elements to these new indices. To determine the new indices for elements with a 0 (F) in the bit, we enumerate these elements as described in the last section. To determine the new indices of elements with a 1 (T) in the bit, we enumerate the elements starting at the top of the vector and subtract these from the length of the vector. Figure 3 shows an example of the `split` operation along with code to implement it.

The `split` operation has a step complexity of $O(1)$; so for d-bit keys, the split radix sort has a step complexity of $O(d)$. If we assume for n keys that the keys are $O(\lg n)$ bits long, a common assumption in models of computation [45], then the algorithm has a step complexity of $O(\lg n)$. Although $O(\lg n)$ is the same asymptotic complexity as existing EREW and CRCW algorithms [1, 11], the algorithm is much simpler and has a significantly smaller constant. Note that since integers, characters, and floating-point numbers can all be sorted with a radix sort, a radix sort suffices for almost all sorting of fixed-length keys required in practice.

The split radix sort is fast in the scan model, but is it fast in practice? After all, our architectural justification claimed that the scan primitives bring the P-RAM models closer to reality. Table 4 compares implementations of the split radix sort and Batcher's bitonic sort [4] on the Connection Machine. We choose the bitonic sort for comparison because it is commonly cited as the most practical parallel sorting algorithm. I have also looked into implementing Cole's sort [11], which is optimal on the P-RAM models, on the Connection Machine. Although never implemented, because of its complexity, it was estimated that it

Flag	=	[T	F	F	T	F	T	T	F]
`enumerate`(Flag)	=	[$\boxed{0}$	1	1	$\boxed{1}$	2	$\boxed{2}$	$\boxed{3}$	4]

A	=	[5	1	3	4	3	9	2	6]
`copy`(A)	=	[5	5	5	5	5	5	5	5]

B	=	[1	1	2	1	1	2	1	1]
`+-distribute`(B)	=	[10	10	10	10	10	10	10	10]

Figure 1: The `enumerate`, `copy`, and `+-distribute` operations. The `enumerate` numbers the flagged elements of a vector, the `copy` copies the first element across a vector, and the `+-distribute` sums the elements of a vector.

`define split-radix-sort`(A, number-of-bits){
  for i **from** 0 **to** (number-of-bits − 1)
    A ← `split`(A, A⟨i⟩)}

A	=	[5	7	3	1	4	2	7	2]
A⟨0⟩	=	[T	T	T	T	F	F	T	F]
A ← `split`(A, A⟨0⟩)	=	[4	2	2	5	7	3	1	7]
A⟨1⟩	=	[F	T	T	F	T	T	F	T]
A ← `split`(A, A⟨1⟩)	=	[4	5	1	2	2	7	3	7]
A⟨2⟩	=	[T	T	F	F	F	T	F	T]
A ← `split`(A, A⟨2⟩)	=	[1	2	2	3	4	5	7	7]

Figure 2: An example of the split radix sort on a vector containing three bit values. The A⟨n⟩ notation signifies extracting the n^{th} bit of each element of the vector A and converting it to a boolean value (T for 1, F for 0). The `split` operation packs `F` (0) elements to the bottom and `T` (1) elements to the top.

```
define split(A, Flags){
        I-down ← enumerate(not(Flags));
        I-up ← n − back-enumerate(Flags) − 1;
        Index ← if Flags then I-up else I-down;
    permute(A, Index)}
```

A	=	[5	7	3	1	4	2	7	2]
Flags	=	[T	T	T	T	F	F	T	F]
I-down	=	[0	0	0	0	$\boxed{0}$	$\boxed{1}$	2	$\boxed{2}$]
I-up	=	[$\boxed{3}$	$\boxed{4}$	$\boxed{5}$	$\boxed{6}$	6	6	$\boxed{7}$	7]
Index	=	[3	4	5	6	0	1	7	2]
permute(A, Index)	=	[4	2	2	5	7	3	1	7]

Figure 3: The `split` operation packs the elements with an `F` in the corresponding flag position to the bottom of a vector, and packs the elements with a `T` to the top of the same vector.

	Split Radix Sort	Bitonic Sort
Theoretical (Bit Serial Circuit) Bit Time	$O(d \lg n)$	$O(d + \lg^2 n)$
Actual (64K processor CM-1) Bit cycles (sorting 16 bits)	20,000	19,000

Table 4: Comparison of the times taken by the split radix sort and the bitonic sort (n keys each with d bits). The constants in the theoretical times are very small for both algorithms. On the Connection Machine, the bitonic sort is implemented in microcode whereas the split radix sort is implemented in macrocode, giving the bitonic sort an edge.

would be at least a factor of 4, and possibly a factor of 10, slower than the other two sorts. The split radix sort is the sort currently supported by the parallel instruction set of the Connection Machine [46].

2.3 Segments and Segmented Scans

In many algorithms it is useful to break the linear ordering of the processors into segments and have a scan operation start again at the beginning of each segment; we call such scan operations, *segmented scans*. Segmented scans take two arguments: a set of values and a set of segment flags. Each flag in the segment flags specifies the start of a new segment (see Figure 4). Segmented scans were first suggested by Schwartz [40] and this paper shows some useful applications of these scans.

The segmented scan operations are useful because they allow algorithms to execute the scans independently over the elements of many sets. This section discusses how they can be used to implement a parallel version of quicksort and how they can be used to represent graphs. This graph representation is then used in a minimum-spanning-tree algorithm.

The segmented scan operations can all be implemented with at most two calls to the two unsegmented primitive scans (see Section 3.4). They can also be implemented directly as described by Schwartz [40].

2.3.1 Example: Quicksort

To illustrate the use of segments, we consider a parallel version of Quicksort. Similar to the standard serial version [24], the parallel version picks one of the keys as a pivot value, splits the keys into three sets—keys lesser, equal and greater than the pivot—and recurses on each set.[4] The algorithm has an expected step complexity of $O(\lg n)$.

The basic intuition of the parallel version is to keep each subset in its own segment, and to pick pivot values and split the keys independently within each segment. The steps required by the sort are outlined as follows:

1. Check if the keys are sorted and exit the routine if they are.

[4] We do not need to recursively sort the keys equal to the pivot, but the algorithm as described below does.

A	=	[5	1	3	4	3	9	2	6]
Sb	=	[T	F	T	F	F	F	T	F]
`seg-+-scan`(A, Sb)	=	[0	5	0	3	7	10	0	2]
`seg-max-scan`(A, Sb)	=	[0	5	0	3	4	4	0	2]

Figure 4: The segmented scan operations restart at the beginning of each segment. The vector Sb contains flags that mark the beginning of the segments.

Key	=	[6.4	9.2	3.4	1.6	8.7	4.1	9.2	3.4]
Segment-Flags	=	[T	F	F	F	F	F	F	F]
Pivots	=	[6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4]
F	=	[=	>	<	<	>	<	>	<]
Key ← `split`(Key, F)	=	[3.4	1.6	4.1	3.4	6.4	9.2	8.7	9.2]
Segment-Flags	=	[T	F	F	F	T	T	F	F]
Pivots	=	[3.4	3.4	3.4	3.4	6.4	9.2	9.2	9.2]
F	=	[=	<	>	=	=	=	<	=]
Key ← `split`(Key, F)	=	[1.6	3.4	3.4	4.1	6.4	8.7	9.2	9.2]
Segment-Flags	=	[T	T	F	T	T	T	T	F]

Figure 5: Parallel quicksort. On each step, within each segment, we distribute the pivot, test whether each element is equal-to, less-than or greater-than the pivot, split into three groups, and generate a new set of segment flags.

Each processor checks to see if the previous processor has a lesser or equal value. We execute an `and-distribute` (Section 2.2) on the result of the check so that each processor knows whether all other processors are in order.

2. Within each segment, pick a pivot and distribute it to the others elements.

   If we pick the first element as a pivot, we can use a segmented version of the `copy` (Section 2.2) operation implemented based on a segmented `max-scan`. The algorithm could also pick a random element by generating a random number in the first element of each segment, moding it with the length of the segment, and picking out the element with a few scans.

3. Within each segment, compare each element with the pivot and split based on the result of the comparison.

   For the split, we can use a version of the `split` operation described in Section 2.2.1 which splits into three sets instead of two, and which is segmented. To implement such a segmented `split`, we can use a segmented versions of the `enumerate` operation (Section 2.2) to number relative to the beginning of each segment, and we can use a segmented version of the `copy` operation to copy the offset of the beginning of each segment across the segment. We then add the offset to the numbers relative to beginning of the segment to generate actual indices to which we permute each element.

4. Within each segment, insert additional segment flags to separate the split values.

   Knowing the pivot value, Each element can determine if its at the beginning of the segment by looking at the previous element.

5. Return to step 1.

Each iteration of this sort requires a constant number of calls to the primitives. If we select pivots randomly within each segment, quicksort is expected to complete in $O(\lg n)$ iterations, and therefore has an expected step complexity of $O(\lg n)$. This version of quicksort has been implemented on the Connection Machine and executes in about twice the time as the split radix sort.

The technique of recursively breaking segments into subsegments and operating independently within each segment can be used for many other divide-and-conquer algorithms [7, 8].

2.3.2 Graphs

An undirected graph can be represented using a segment for each vertex and an element position within a segment for each edge of the vertex. Since each edge is incident on two vertices, it appears in two segments. The actual values kept in the elements of the segmented vector are pointers to the other end of the edge (see Figure 6). To include weights on the edges of the graphs, we can use an additional vector that contains the weights of the edges.

By using segmented operations to operate over the edges of each vertex, the step complexity of many useful operations on graphs can be reduced. For example, for n vertices, the complexity of each vertex summing a value from all neighbors, is reduced from $O(\lg n)$ in the P-RAM models to $O(1)$ in the scan model. Such neighbor summing can be executed by distributing the value from each vertex over its edges using a segmented `copy` operation, permuting these values using the cross-pointers, and summing the values on the edges back into the vertices using a segmented `+-distribute` operation. Elsewhere [7] we show that by keeping trees in a particular form, we can similarly reduce the step complexity of many tree operations on trees with n vertices by $O(\lg n)$.

For this segmented graph representation to be useful, either there must be an efficient routine to generate the representation from another representation, or it must be possible to do all manipulations on graphs using the segmented graph representation. A graph can be converted from most other representations into the segmented graph representation by creating two elements per edge (one for each end) and sorting the edges according to their vertex number. The split radix sort (Section 2.2.1) can be used since the vertex numbers are all integers less than n. The sort places all edges that belong to the same vertex in a contiguous segment. We suggest, however, that in the scan model graphs always be kept in the segmented graph representation.

2.3.3 Minimum Spanning Tree

This section describes a probabilistic minimum-spanning-tree (MST) algorithm. For n vertices and m edges, it has a step complexity of $O(\lg n)$. The best algorithm known for the EREW P-RAM model requires $O(\lg^2 n)$ time [23, 39]. The best algorithm known for the CRCW P-RAM model requires $O(\lg n)$ time [3], but this algorithms requires that the generic CRCW P-RAM model be extended so that if several processors write to the same location, either the value from the lowest numbered processor is written, or the minimum value is written.

All these algorithms are based on the algorithm of Sollin [5], which is similar to the algorithm of Borůvka [9]. The algorithms start with a forest of trees in which each tree is a single vertex. These trees are merged during the algorithm, and the algorithm terminates when a single tree remains. At each step, every tree T finds its minimum-weight edge joining a vertex in T to a vertex of a distinct tree T'. Trees connected by one of these edges merge. To reduce the forest to a single tree, $O(\lg n)$ such steps are required.

In the EREW P-RAM algorithm, each step requires $\Omega(\lg n)$ time because finding the minimum edge in a tree and distributing connectivity information over merging trees might require $\Omega(\lg n)$ time. In the extended CRCW P-RAM model, each step only requires constant time because each minimum edge can be found with a single write operation. In our algorithm, we keep the graph in the graph representation discussed in Section 2.3.2 so that we can use the `min-distribute` (Section 2.2) operation to find the minimum edge for each tree and the `copy` operation to distribute connectivity information among merging trees with a constant number of calls to the primitives. The only complication is maintaining the representation when merging stars.

As with the Shiloach and Vishkin CRCW P-RAM algorithm [43], trees are selected for merging by forming stars. We define a star as a set of vertices within a graph with one of the set marked as the parent, the others marked as children, and an edge that leads from each child vertex to its parent vertex[5]. A graph might contain many stars. The *star-merge* operation takes a graph with a set of disjoint stars, and returns a graph with each star merged into a single vertex. Figure 7 shows an example of a star-merge for a graph with a single star.

The minimum-spanning-tree algorithm consists of repeatedly finding starts and merging them. To find stars, each vertex flips a coin to decide whether they are a child or parent. All children find their minimum edge (using a `min-distribute`), and all such edges that are connected to a parent are marked as *star edges*. Since, on average, half the trees are children

[5]This definition of a star is slightly different from the definition of Shiloach and Vishkin [43].

and half of the trees on the other end of the minimum edge of a child are parents, 1/4 of the trees are merged on each star-merge step. This random mate technique is similar to the method discussed by Miller and Reif [33]. Since on average 1/4 of the trees are deleted on each step, $O(\lg n)$ steps are required to reduce the forest to a single tree.

We now describe how a star-merge operation can be implemented in the scan model, such that for m edges, the operation has a step complexity of $O(1)$. The input to the star-merge operation is a graph in the segmented graph representation, with two additional vectors: one contains flags that mark every star edge, and the other contains a flag that marks every parent. To implement a star-merge and maintain the segmented graph representation, each child segment must be moved into its parent segment. The technique used for this rearrangement can be partitioned into four steps: (1) each parent opens enough space in its segment to fit its children, (2) the children are permuted into this space, (3) the cross-pointers vector is updated to reflect the change in structure of the graph, and (4) edges which point within a segment are deleted, therefore deleting edges that point within a tree.

(1) To open space in the parent segments, each child passes its length (number of edges) across its star edge to its parent, so each parent knows how much space it needs to open up for each of its children. Let us call the vector that contains the needed space of each child the *needed-space* vector. A 1 is placed in all the nonstar edges of this vector. We can now use a segmented `+-distribute` on the *needed-space* vector to determine the new size of each parent and a `+-scan` to allocate this new space for each parent (such allocation is discussed in more detail in Section 2.4). We also execute a segmented `+-scan` on the *needed-space* vector to determine the offset of each child within its parent segment and the new position of each non star edge of the parent. We call this vector the *child-offset* vector.

(2) We now need to permute the children into the parent segments. To determine the new position of the edges in the child vertices, we permute the *child-offset* back to each child and distribute it across the edges of the child. Each child adds its index to this offset giving each child edge a unique address within the segment of its parent. We now permute all the edges, children and parents, to their new positions. (3) To update the pointers, we simply pass the new position of each end of an edge to the other end of the edge. (4) To delete edges that point within a segment, we check if each edge points within the segment by distributing the ends of the segment, pack all elements that point outside each segment deleting elements pointing within each segment, and update the pointers again.

2.4 Allocating

This section illustrates another use of the scan operations. Consider the problem of given a set of processors each with an integer, allocating that integer number of new processors to each initial processor. Such allocation is necessary in the parallel line-drawing routine described in Section 2.4.1. In the line-drawing routine, each line calculates the number of pixels in the line and dynamically allocates a processor for each pixel. Allocating new elements is also useful for the branching part of many branch-and-bound algorithms. Consider, for example, a brute force chess-playing algorithm that executes a fixed-depth search of possible moves to determine the best next move.[6] We can execute the algorithms in parallel by placing each possible move in a separate processor. Since the algorithm dynamically decides how many next moves to generate, depending on the position, we need to dynamically allocate new elements. In Section 2.5 we discuss the bounding part of branch-and-bound algorithms.

Defined more formally, allocation is the task of, given a vector of integers A with elements a_i and length l, creating a new vector B of length:

$$L = \sum_{i=0}^{l-1} a_i \tag{1}$$

with a_i elements of B assigned to each position i of A. By assigned to, we mean that there must be some method for distributing a value at position i of a vector to the a_i elements which are assigned to that position. Since there is a one-to-one correspondence between elements of a vector and processors, the original vector requires l processors and the new vector requires L processors. Typically, an algorithm does not operate on the two vectors at the same time, so the processors can overlap.

Allocation can be implemented by assigning a contiguous segment of elements to each position i of A. To allocate segments we execute a `+-scan` on the vector A returning a pointer to the start of each segment (see Figure 8). We can then generate the appropriate segment flags by permuting a flag to the index specified by the pointer. To distribute values from each position i to its segment, we permute the values to the beginning of the segments and use a segmented `copy` operation to copy the values across the segment. Allocation and

[6]This is how many chess playing algorithms work [6]. The search is called an *minimax* search since it alternates moves between the two players, trying to minimize the benefit of one player and maximize the benefit of the other.

distribution each require $O(1)$ steps on the scan model. Allocation requires $O(\lg n)$ steps on a EREW P-RAM and $O(\lg n / \lg \lg n)$ steps on a CREW P-RAM (this is based on the prefix sum routine of Cole and Vishkin [13]).

When an algorithm allocates processors, the number of processors required is usually determined dynamically and will depend on the data. To account for this, we must do one of three things: assume an infinite number of processors, put a bound on the number of elements that can be allocated, or start simulating multiple elements on each processor. The first is not practical, and the second restricting. Section 2.5 discusses the third.

2.4.1 Example: Line Drawing

As a concrete example of how allocation is used, consider line drawing. Line drawing is the problem of, given a set of pairs of points (each point is an (x, y) pair), generating all the locations of pixels that lie between one of the pairs of points. Figure 9 illustrates an example. The routine we discuss returns a vector of (x, y) pairs that specify the position of each pixel along the line. It generates the same set of pixels as generated by the simple digital differential analyzer (DDA) serial technique [34].

The basic idea of the routine is for each line to allocate a processor for each pixel in the line, and then for each allocated pixel to determine, in parallel, its final position in the grid. To allocate a processor for each pixel, each line must first determine the number of pixels in the line. This number can be calculated by taking the maximum of the x and y differences of the line's endpoints. Each line now allocates a segment of processors for its pixels, and distributes its endpoints across the segment as described earlier. We now have one processor for each pixel and one segment for each line. We can view the position of a processor in its segment as the position of a pixel in its line. Based on the endpoints of the line and the position in the line (determined with a `+-scan`), each pixel can determine its final (x, y) location in the grid. To actually place the points on a grid, we would need to permute a flag to a position based on the location of the point. In general, this will require the simplest form of concurrent-write (one of the values gets written) since a pixel might appear in more than one line.

This routine has a step complexity of $O(1)$ and requires as many processors as pixels in the lines. The routine has been implemented on the Connection Machine, has been extended to render solid objects by Salem, and is part of a rendering package for the

Connection Machine [38].

2.5 Load Balancing

Up to now this paper has assumed that a P-RAM always has as many processors as elements in the data vectors. This section considers simulating multiple elements on each processor. Such simulation is important for two reasons. First, from a practical point of view, real machines have a fixed number of processors but problem sizes vary: we would rather not restrict ourselves to fixed, and perhaps small, sized problems. Second, from both a practical and theoretical point of view, by placing multiple elements on each processor, an algorithm can more efficiently utilize the processors and can greatly reduce the processor-step complexity (see Table 5). Figure 10 discusses how to simulate the various vector operations discussed in Section 2.1 on vectors with more elements than processors.

When simulating multiple elements on each processor, it is important to keep the number of elements on the processors balanced. We call such balancing, *load-balancing*. Load-balancing is important when data elements drop out during the execution of an algorithm since this might leave the remaining elements unbalanced. There are three common reasons why elements might drop out. First, some elements might have completed their desired calculations. For example, in the quicksort algorithm described in Section 2.3.1, segments which are already sorted might drop out. Second, the algorithm might be subselecting elements. Subselection is used in the halving merge algorithm discussed in Section 2.5.1. Third, an algorithm might be pruning some sort of search. Pruning might be used in the bounding part of branch-and-bound algorithms such as the chess-playing algorithm we mentioned in Section 2.4. In all three cases, when the elements drop out, the number of elements left on each processor might be unbalanced.

For m remaining elements, load balancing can be implemented by enumerating the remaining elements, permuting them into a vector of length m, and assigning each processor to m/p elements of the new vector (see Figure 11). We call the operation of packing elements into a smaller vector, the `pack` operation. A processor can determine how many and which elements it is responsible for simply by knowing its processor number and m; m can be distributed to all the processors with a `copy`. On an EREW P-RAM, load balancing requires $O(n/p + \lg n)$ steps.

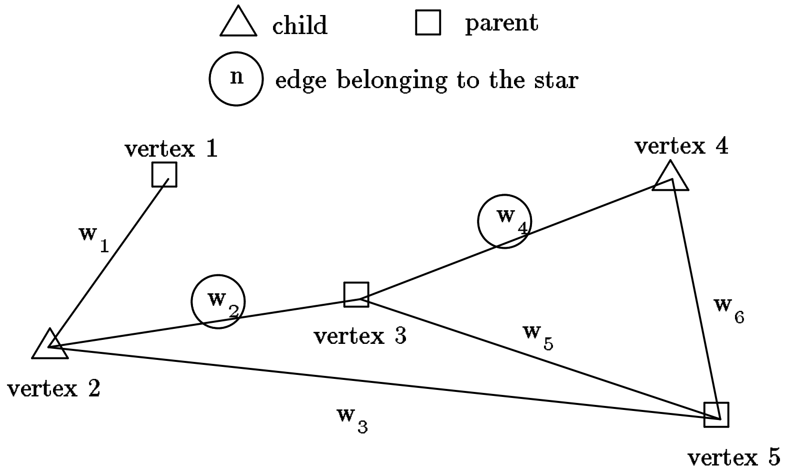
Sorry figure has been lost

$$
\begin{array}{lcl|ccc|ccc|cc|ccc}
\text{Index} & = & [0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11] \\
\text{vertex} & = & [1 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 5] \\
\text{segment-descriptor} & = & [\mathtt{T} & \mathtt{T} & \mathtt{F} & \mathtt{F} & \mathtt{T} & \mathtt{F} & \mathtt{F} & \mathtt{T} & \mathtt{F} & \mathtt{T} & \mathtt{F} & \mathtt{F}] \\
\text{cross-pointers} & = & [1 & 0 & 4 & \boxed{9} & 2 & 7 & 10 & 5 & 11 & \boxed{3} & 6 & 8] \\
\text{weights} & = & [w_1 & w_1 & w_2 & w_3 & w_2 & w_4 & w_5 & w_4 & w_6 & w_3 & w_5 & w_6]
\end{array}
$$

Figure 6: An example of the undirected segmented graph representation. Each pointer points to the other end of the edge. So, for example, edge w_3 in vertex 2 contains a pointer (in this case 9) to its other end in vertex 5. The segment-descriptor partitions the processors into vertices.

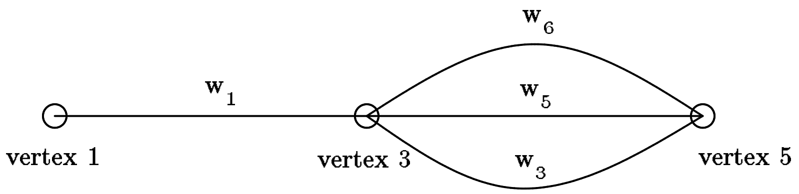
	Processors	Steps	Processor-Step
Halving Merge	$O(n)$	$O(\lg n)$	$O(n \lg n)$
	$O(n/\lg n)$	$O(\lg n)$	$O(n)$
List Ranking [12]	$O(n)$	$O(\lg n)$	$O(n \lg n)$
	$O(n/\lg n)$	$O(\lg n)$	$O(n)$
Tree Contraction [18]	$O(n)$	$O(\lg n)$	$O(n \lg n)$
	$O(n/\lg n)$	$O(\lg n)$	$O(n)$

Table 5: The processor-step complexity of many algorithms can be reduced by using fewer processors and assigning many elements to each processor.



Before Star-Merge

Index	=	[0	1	2	3	4	5	6	7	8	9	10	11]
segment-descriptor	=	[T	T	F	F	T	F	F	T	F	T	F	F]
weights	=	[w_1	w_1	w_2	w_3	w_2	w_4	w_5	w_4	w_6	w_3	w_5	w_6]
Star-Edge	=	[F	F	T	F	T	T	F	T	F	F	F	F]
Parent	=	[T	F			T			F		T]		



After Star-Merge

Index	=	[0	1	2	3	4	5	6	7]
segment-descriptor	=	[T	T	F	F	F	T	F	F]
weights	=	[w_1	w_1	w_3	w_5	w_6	w_3	w_5	w_6]
cross-pointers	=	[1	0	5	6	7	2	3	4]

Figure 7: An example of star merging. In this example, two parents have no children (1 and 5) and the third (3) has two children (2 and 4). The second diagram shows the graph after the star is merged.

$$\begin{array}{ll} \text{V} & = [v_1 \quad v_2 \quad v_3] \\ \text{A} & = [4 \quad 1 \quad 3] \\ \text{Hpointers} \leftarrow +\texttt{-scan}(\text{A}) & = [0 \quad 4 \quad 5] \end{array}$$

$$\begin{array}{ll} \text{Segment-flag} & = [1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0] \\ \texttt{distribute}(\text{V, Hpointers}) & = [v_1 \quad v_1 \quad v_1 \quad v_1 \quad v_2 \quad v_3 \quad v_3 \quad v_3] \end{array}$$

Figure 8: Processor Allocation. The vector A specifies how many new elements each position needs. We can allocate a segment to each position by applying a +`-scan` to A and using the result as pointers to the beginning of each segment. We can then distribute the values of V to the new elements with a permute to the beginning of the segment and a segmented `copy` across the segment.

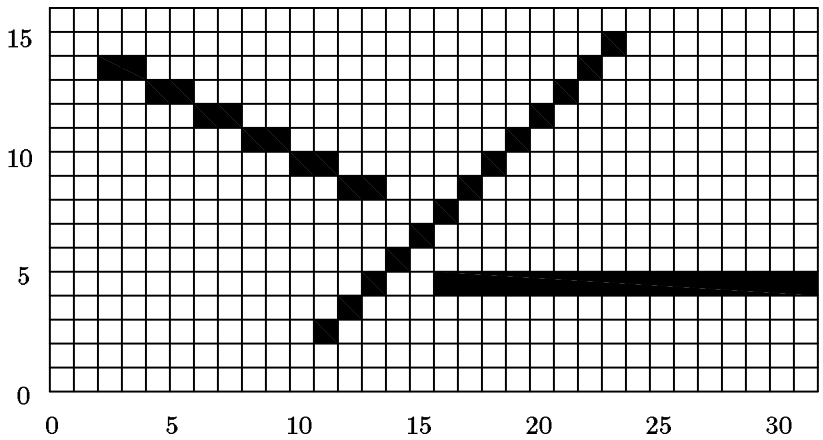


Figure 9: The pixels generated by a line drawing routine. In this example the endpoints are (11, 2)–(23, 14), (2, 13)–(13,8), and (16, 4)–(31, 4). The algorithm allocates 12, 11 and 16 pixels respectively for the three lines.

$$[\underbrace{4 \quad 7 \quad 1}_{\text{processor } 0} \quad \underbrace{0 \quad 5 \quad 2}_{\text{processor } 1} \quad \underbrace{6 \quad 4 \quad 8}_{\text{processor } 2} \quad \underbrace{1 \quad 9 \quad 5}_{\text{processor } 3}]$$

$$\begin{array}{lcllll} \text{Sum} & = & [12 & 7 & 18 & 15] \\ +\texttt{-scan}(\text{Sum}) & = & [0 & 12 & 19 & 37] \end{array}$$

$$[\underbrace{0 \quad 4 \quad 11}_{\text{processor } 0} \quad \underbrace{12 \quad 12 \quad 17}_{\text{processor } 1} \quad \underbrace{19 \quad 25 \quad 29}_{\text{processor } 2} \quad \underbrace{37 \quad 38 \quad 47}_{\text{processor } 3}]$$

Figure 10: When operating on vectors with more elements than processors (*long vectors*), each processor is assigned to a contiguous block of elements. To execute an arithmetic operation or the `permute` operation on a long vector, each processor loops over the element positions for which it is responsible and executes the operation. To execute a scan, each processor sums its elements, executes a scan across processors, and uses the result as an offset to scan within processors.

$$\begin{array}{lcl} F & = & [T \quad F \quad F \quad F \quad T \quad T \quad F \quad T \quad T \quad T \quad T \quad T] \\ A & = & [\underbrace{a_0 \quad a_1 \quad a_2}_{\text{processor } 0} \quad \underbrace{a_3 \quad a_4 \quad a_5}_{\text{processor } 1} \quad \underbrace{a_6 \quad a_7 \quad a_8}_{\text{processor } 2} \quad \underbrace{a_9 \quad a_{10} \quad a_{11}}_{\text{processor } 3}] \end{array}$$

$$A = [\underbrace{a_0 \quad a_4}_{\text{proc } 0} \quad \underbrace{a_5 \quad a_7}_{\text{proc } 1} \quad \underbrace{a_8 \quad a_9}_{\text{proc } 2} \quad \underbrace{a_{10} \quad a_{11}}_{\text{proc } 3}]$$

Figure 11: Load Balancing. In load balancing, certain marked elements are dropped and the remaining elements need to be balanced across the processors. Load balancing can be executed by packing the remaining elements into a smaller vector using an `enumerate` and a `permute`, and assigning each processor to a smaller block.

2.5.1 Example: Halving Merge

To illustrate the importance of simulating multiple elements on each processor and load balancing, this section describes an algorithm for merging two sorted vectors.[7] We call the algorithm, the *halving merge*. When applied to vectors of length n and m ($n \geq m$) on the scan model with p processors, the halving merge algorithm has an step complexity of $O(n/p + \lg n)$. When the $p < n/\lg n$, the algorithm is optimal. Although the split radix sort and the quicksort algorithms are variations of well-known algorithms translated to a new model, the merging algorithm described here is original. The merging algorithm of Shiloach and Vishkin for the CRCW P-RAM model [17, 42] has the same complexity but is quite different. Their algorithm is not recursive.

The basic idea of the halving merge algorithm is to extract the odd-indexed elements from each of the two vectors by packing them into smaller vectors, to recursively merge the half-length vectors, and then to use the result of the halving merge to determine the positions of the even-indexed elements. The number of elements halves on each recursive call, and the recursion completes when one of the merge vectors contains a single element. We call the operation of taking the result of the recursive merge on the odd-indexed elements and using it to determine the position of the even-indexed elements, *even-insertion*. We first analyze the complexity of the halving merge assuming that the even-insertion requires a constant number of scan and permute operations, and then discuss the algorithm in more detail.

The complexity of the algorithm is calculated as follows. Since the number of elements halves at each level, there are at most $\lg n$ levels and at level i, $n/2^i$ elements must be merged. With p processors, if we load balance, the most elements any processor is responsible for is

$$\lceil n/2^i p \rceil \ . \tag{2}$$

If the even-insertion requires a constant number of calls to the primitives per element, level i has a step complexity of

$$O(\lceil n/2^i p \rceil) \ . \tag{3}$$

[7] In this algorithm, the elements drop out in a very regular fashion and the remaining elements could be balanced without the scan primitives. Load balancing using scans, however, simplifies the description.

The total step complexity is therefore

$$\begin{aligned} O\left(\sum_{i=0}^{\lg n-1}\left\lceil\frac{n}{2^i p}\right\rceil\right) &= O\left(\sum_{i=0}^{\lg n-1}\left(\frac{n}{2^i p}+1\right)\right) \\ &= O\left(n/p+\lg n\right) \ . \end{aligned} \tag{4}$$

We now discuss the algorithm in more detail. Picking every other element before calling the algorithm recursively can be implemented by marking the odd-indexed elements and packing them (load balancing them). After the recursive call returns, the even-insertion is executed as follows. We expand the merged odd-indexed vector by a factor of two by placing each unmerged even-indexed element directly after the element it originally followed (see Figure 12). We call this vector the *near-merge* vector. The *near-merge* vector has an interesting property: elements can only be out of order by single nonoverlapping rotations. An element might appear before a block of elements it belongs after. We call such an element a block-head. A near-merge vector can be converted into a true merged vector by moving the block-head to the end of the block and sliding the other elements down by one: rotating the block by one. The rotation of the blocks can be implemented with two scans and two arithmetic operations:

```
define fix-near-merge(near-merge){
    head-copy ← max(max-scan(near-merge), near-merge)
    result ← min(min-backscan(near-merge), head-copy)}
```

The first step moves the block-head to the end of the block, and the second step shifts the rest of the block down by one. The even-insertion therefore requires a constant number of calls to the vector operations.

To place the even-indexed elements following the odd-indexed elements after returning from the recursive call, we must somehow know the original position of each merged odd-indexed element. To specify these positions, the merge routine could instead of returning the actual merged values, return a vector of flags: each `F` flag represents an element of A and each `T` flag represents an element of B. For example:

$$\begin{array}{lcllll} A' & = & [1 & 10 & 15] \\ B' & = & [3 & 9 & 23] \end{array}$$

$$
\begin{array}{lcllllll}
\mathrm{A} & = & [1 & 7 & 10 & 13 & 15 & 20] \\
\mathrm{B} & = & [3 & 4 & 9 & 22 & 23 & 26]
\end{array}
$$

$$
\begin{array}{lclll}
\mathrm{A}' & = & [1 & 10 & 15] \\
\mathrm{B}' & = & [3 & 9 & 23]
\end{array}
$$

$$
\mathrm{merge}(\mathrm{A}', \mathrm{B}') = [1 \quad 3 \quad 9 \quad 10 \quad 15 \quad 23]
$$

$$
\begin{array}{lcl}
\text{near-merge} & = & [1 \quad \boxed{7 \quad 3 \quad 4} \quad 9 \quad \boxed{22 \quad 10 \quad 13 \quad 15 \quad 20} \quad 23 \quad 26] \\
\text{result} & = & [1 \quad 3 \quad 4 \quad 7 \quad 9 \quad 10 \quad 13 \quad 15 \quad 20 \quad 22 \quad 23 \quad 26]
\end{array}
$$

Figure 12: The halving merge involves selecting the odd-indexed elements of each vector to be merged, recursively merging these elements and then using the result to merge the even-indexed elements (even-insertion). To execute the even-insertion, we place the even-indexed elements in the merged odd-indexed elements after their original predecessor. This vector, the *near-merge* vector, is almost sorted. As shown in the figure, nonoverlapping blocks might need to be rotated: the first element moved to the end.

$$\text{halving-merge}(A', B') = [\texttt{F} \quad \texttt{T} \quad \texttt{T} \quad \texttt{F} \quad \texttt{F} \quad \texttt{T}]$$

which corresponds to the merged values:

$$[1 \quad 3 \quad 9 \quad 10 \quad 15 \quad 23]$$

The vector of flags—henceforth the *merge-flag* vector—both uniquely specifies how the elements should be merged and specifies in which position each element belongs.

3 Implementation

This section describes a circuit that implements two scan primitives, an integer version of the `+-scan` and `max-scan`, and describes how the other scans used in this paper can be simulated with the two primitives. From a theoretical orientation, efficient circuits for implementing scan primitives have been discussed elsewhere [28, 15]. This section therefore concentrates on a practical implementation, described at the logic level, and discusses how this implementation could fit into an actual machine. Elsewhere we have shown [7] that some of the other scan operations, such as the segmented scan operations, can be implemented directly with little additional hardware.

Although the discussion suggests a separate circuit (set of chips) to implement the scan operations, wires and chips of a scan circuit might be shared with other circuitry. The scan implementation on the Connection Machine, for example, shares the wires with the router and requires no additional hardware.

3.1 Tree Scan

Before describing details on how a circuit is implemented, we review a standard, general technique for implementing the scan operation on a balanced binary tree for any binary associative scan operator $\oplus$. The technique consists of two sweeps of the tree, an up sweep and a down sweep, and requires $2 \lg n$ steps. Figure 13 shows an example. The values to be scanned start at the leaves of the tree. On the up sweep, each unit executes $\oplus$ on its two children units and passes the sum to its parent. Each unit also keeps a copy of the value from the left child in its memory. On the down sweep, each unit passes to its left child the value from its parent and passes to its right child $\oplus$ applied to its parent and the value stored in the memory (this value originally came from the left child). After the down sweep, the values at the leaves are the results of a scan.

If the scan operator $\oplus$ can be executed with a single pass over the bits of its operand, such as integer addition and integer maximum, the tree algorithm can be bit pipelined. Bit pipelining involves passing the operands one bit at a time up the tree so that when the second level is working on bit n the first level works on bit $n+1$. Such bit pipelining can greatly reduce the hardware necessary to implement the scan operations since only single bit logic is required at each unit.

As an example of how bit pipelining works, we consider a bit-pipelined version of `+-scan` for n, m bit integers. This bit-pipelined scan starts by passing the least significant bit of each value into the leaf units of the tree. Each unit now performs a single bit addition on its two input bits, stores the carry bit in a flag, propagates the sum bit to its parent in the next layer of units, and stores the bit from the left child in an m bit memory on the unit. On the second step, the scan passes the second bit of each value into the leaf units of the tree while it propagates the least significant bit of the sums on the first layer, to the second layer. In general, on the i^{th} step, at the j^{th} layer (counting from the leaves), the $(i-j)^{th}$ bit of the sum of a unit (counting from the least significant bit) gets propagated to its parent. After $m + \lg n$ steps, the up sweep is completed. Using a similar method, the down-sweep is also calculated in $m + \lg n$ steps. The total number of steps is therefore $2(m + \lg n)$. The down sweep can actually start as soon as the first bit of the up sweep reaches the top, reducing the number of steps to $m + 2\lg n$.

3.2 Hardware Implementation of Tree Scan

We now discuss in more detail the hardware needed to implement the bit-pipelined tree scan for the two primitive scan operations `+-scan` and `max-scan`. Figure 14 shows an implementation of a unit of the binary tree. Each unit consist of two identical state machines, a variable-length shift register and a one bit register. The control for a unit consists of a clock, a clear signal, and an operation specification, which specifies whether to execute a `+-scan` or a `max-scan`. The control signals are identical on all units. The units are connected in a balanced binary tree, as shown in Figure 13, with two single bit unidirectional wires along every edge.

The *shift register* acts as a first in first out buffer (FIFO), with bits entered on one end

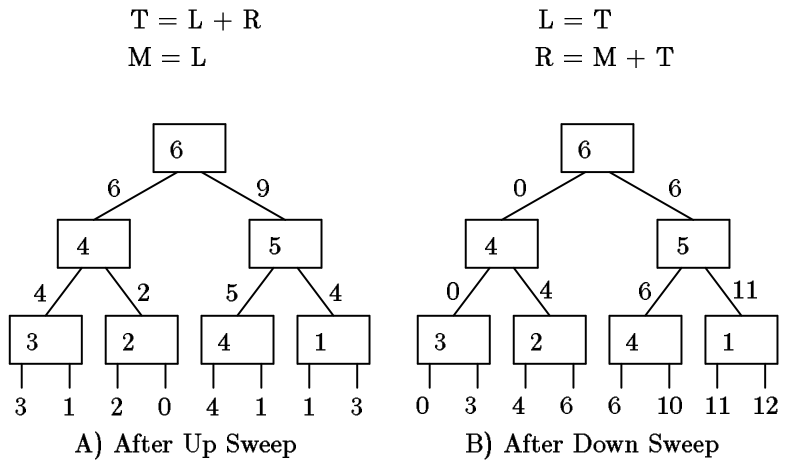


Figure 13: Parallel scan on a tree using the operator "+". The number in the block is the number being stored in the memory on the up sweep.

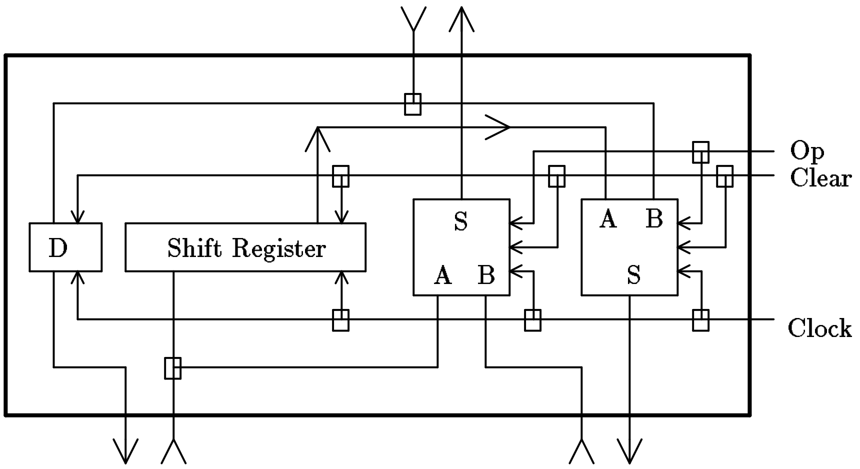


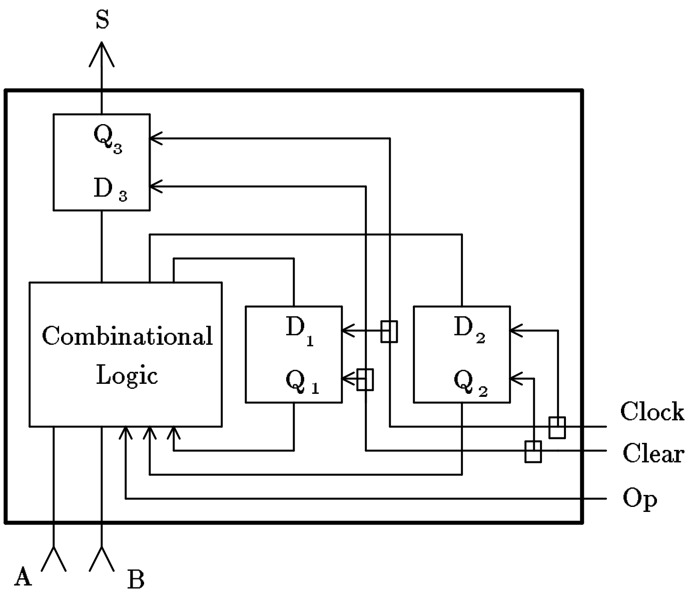
Figure 14: Diagram of a unit needed to implement the tree algorithm. It consists of a shift register (which acts as a first in first out buffer), a one bit register (a D type flip-flop), and two identical sum state machines. These units are arranged in a tree as shown in Figure 13. Each wire is a single bit wire.

and removed from the other. One bit is shifted on each clock signal. The length of the register depends on the depth of the unit in the tree. A unit at level i from the top needs a register of length $2i$ bits. The maximum length is therefore $2 \lg n$ bits. The length of the shift register can either be hardwired into each unit, in which case different levels of the tree would require different units, or could be controlled by some logic, in which case all units could be identical, but the level number would need to be stored on each unit.

The *sum state machine* consists of three bits of state and a five input, three output combinational logic circuit. Figure 15 shows its basic layout and the required logic. The *Op* control specifies whether to execute a `+-scan` or a `max-scan`. Two bits of state are needed for the `max-scan` to keep track of whether A is greater, equal or lesser than B (if Q_1 is set, A is greater, if Q_2 is set, B is greater). The `+-scan` only uses one bit of state to store the carry flag (Q_1). The third bit of state is used to hold the output value (S) for one clock cycle.

To execute a scan on a tree of such units, we reset the state machines in all units with the *clear* signal and set the *Op* signal to execute either a `+-scan` or `max-scan`. We must tie the parent input of the root unit low (0). We then simply insert one bit of the operand at the leaves on each clock cycle. In a `max-scan` the bits are inserted starting at the most significant bit, and in a `+-scan` the bits are inserted starting at the least significant bit. After $2 \lg n$ steps, the result will start returning at the leaves one bit on each clock cycle. We do not even need to change anything when going from the up sweep to the down sweep: when the values reach the root, they are automatically reflected back down since the shift register at the root has length 0. The total hardware needed for scanning n values is $n-1$ shift registers and $2(n-1)$ sum state machines. The units are simple so it should be easy to place many on a chip.

Perhaps more importantly than the simplicity of each unit is the fact that the units are organized in a tree. The tree organization has two important practical properties. First, only two wires are needed to leave every branch of the tree. So, for example, if there are several processors per chip, only a pair of wires are needed to leave that chip, and if there are many processors on a board, only a pair of wires are needed to leave the board. Second, a tree circuit is much easier to synchronize than other structures such as grids, hypercubes or butterfly networks. This is because the same tree used for the scans can be used for



$$
\begin{array}{rcllcl}
 & & \underline{\text{For } \texttt{max-scan}} & & \underline{\text{For } \texttt{+-scan}} \\
S & = & \mathrm{Op}(B\overline{Q_1} + A\overline{Q_2}) & + & \overline{\mathrm{Op}}(A \oplus B \oplus Q_1) \\
D_1 & = & \mathrm{Op}(Q_1 + \overline{B}A\overline{Q_2}) & + & \overline{\mathrm{Op}}(AB + AQ_1 + BQ_1) \\
D_2 & = & \mathrm{Op}(Q_2 + \overline{A}B\overline{Q_1}) & &
\end{array}
$$

Figure 15: Diagram of the sum state machine. It consists of three d-type flip-flops and some combinational logic. If the signal *Op* is true, the circuit executes a `max-scan`. If the signal *Op* is false, the circuit executes a `+-scan`. In the logic equations, the symbol $\oplus$ is an exclusive-or, and the symbol + is an inclusive-or. This state machine fits into a unit as shown in Figure 14.

clock distribution. Such a clock distribution gets rid of the clock skew problem[8] and makes it relatively easy to run the circuit extremely fast.

3.3 An Example System

We now consider an example system to show how the scan circuit might be applied in practice. We consider a 4096 processor parallel computer with 64 processors on each board and 64 boards per machine. To implement the scan primitives on such a machine, we could use a single chip on each board that has 64 inputs and 1 output and acts as 6 levels of the tree. Such a chip would require 126 sum state machines and 63 shift registers—such a chip is quite easy to build with today's technology. We could use one more of these chips to combine the pair of wires from each of the 64 boards.

If the clock period is 100 nanoseconds, a scan on a 32 bit field would require 5 microseconds. This time is considerably faster than the routing time of existing parallel computers such as the BBN Butterfly or the Thinking Machines Connection Machine. With a more aggressive clock such as the 10 nanoseconds clock being looked at by BBN for the Monarch[9] [2], this time would be reduced to .5 microseconds—twice as fast as the best case global access time expected on the Monarch.

In most existing and proposed tightly connected parallel computers [22, 36, 2, 41], the cost of the communication network is between 1/3 and 1/2 the cost of the computer. It is unlikely that the suggested scan network will be more than 1% of the cost of a computer.

3.4 Simulating All Scans with a +-Scan and Max-Scan

All the scans discussed in this paper, including the segmented versions, can be implemented just two scans: integer versions of the `+-scan` and `max-scan`. This section discusses the implementation. The implementation requires access to the bit representation of the numbers.

An integer `min-scan` can be implemented by inverting the source, executing a `max-scan`, and inverting the result. A floating-point `max-scan` and `min-scan` can be implemented by

[8] When there are many synchronous elements in a system, the small propagation time differences in different paths when distributing the clock signals can cause significant clock time differences at the elements.

[9] Because of the tree structure, it would actually be much easier to run a clock at 10 nanoseconds on a scan network than it is for the communication network of the Monarch.

flipping the exponent and significant if the sign bit is set, executing the signed version, and flipping the exponent and significant of the result back based on the sign bit. The `or-scan` and `and-scan` can be implemented with a 1-bit `max-scan` and `min-scan` respectively. The implementation of the floating-point `+-scan` is described elsewhere [7].

A segmented `max-scan` is implemented by first enumerating the segment bits, appending the result (plus 1 in positions where the flag is set) to the original numbers, executing an unsegmented `max-scan`, and removing the appended bits (see Figure 16). A segmented `+-scan` is implemented by executing an unsegmented `+-scan`, copying the first element in each segment across the segment using a `copy` (can be implemented with a segmented `max-scan`), and subtracting this element.

The backward scans can be implemented by simply reading the vector into the processors in reverse order.

4 Conclusions

This paper described a study of the effects of adding two scan primitives as unit-time primitives to the P-RAM models. The study shows that these scan primitives take no longer to execute than parallel memory references, both in practice and in theory, but yet can improve the asymptotic running time, and the description of many algorithms. We wonder whether there might be other primitives that can be cheaply implemented on a parallel architecture. One such primitive might be a merge primitive that merges two sorted vectors. As shown by Batcher [4], this can be executed in a single pass of an Omega network.

We hope that this paper will prompt researchers to question the validity of the pure PRAM models when designing and analyzing algorithms—especially when it comes down to trying to remove $O(\lg \lg n)$, or even $\lg n$ factors off of the running time of an algorithm.

Acknowledgments

Many thanks to Charles Leiserson for his ideas, support and enthusiasm, Abhiram Ranade for first pointing out to me that scans can be bit pipelined, and to Andrew Goldberg, Danny Hillis, Tom Knight, John Rose and Guy Steele for their helpful comments.

References

[1] M. Ajtai, J. Komlos, and E. Szemeredi. An $O(n \lg n)$ sorting network. In *Proceedings ACM Symposium on Theory of Computing*, pages 1–9, April 1983.

[2] D. C. Allen. The BBN multiprocessors: Butterfly and Monarch. In *Proceedings Princeton Conference on Supercomputers and Stellar Dynamics*, June 1986.

[3] Baruch Awerbuch and Yossi Shiloach. New connectivity and MSF algorithms for Ultracomputer and PRAM. In *Proceedings ACM Symposium on Theory of Computing*, pages 175–179, 1985.

[4] Kenneth E. Batcher. Sorting networks and their applications. In *AFIPS Spring Joint Computer Conference*, pages 307–314, 1968.

[5] C. Berge and A. Ghouila-Houri. *Programming, Games, and Transportation Networks.* John Wiley, New York, 1965.

[6] H. J. Berliner. A chronology of computer chess and its literature. *Artificial Intelligence*, 10(2), 1978.

[7] Guy E. Blelloch. *Scan Primitives and Parallel Vector Models.* PhD thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, November 1988.

[8] Guy E. Blelloch and James J. Little. Parallel solutions to geometric problems on the Scan Model of computation. In *Proceedings International Conference on Parallel Processing*, pages Vol 3: 218–222, August 1988.

[9] O. Borůvka. O jistém problén minimálím. *Práca Moravské Přírodovědecké Společnosti*, (3):37–58, 1926. (In Czech.).

[10] R. P. Brent and H. T. Kung. The chip complexity of binary arithmetic. In *Proceedings ACM Symposium on Theory of Computing*, pages 190–200, 1980.

[11] Richard Cole. Parallel merge sort. In *Proceedings Symposium on Foundations of Computer Science*, pages 511–516, October 1986.

[12] Richard Cole and Uzi Vishkin. Approximate scheduling, exact scheduling, and applications to parallel algorithms. In *Proceedings Symposium on Foundations of Computer Science*, pages 478–491, 1986.

[13] Richard Cole and Uzi Vishkin. *Faster Optimal Parallel Prefix Sums and List Ranking*. Technical Report Ultracomputer Note 117, New York University, February 1987.

[14] C. C. Elgot and A. Robinson. Random access stored program machines. *Journal of the Association of Computer Machinery*, 11(4):365–399, 1964.

[15] Faith E. Fich. New bounds for parallel prefix circuits. In *Proceedings ACM Symposium on Theory of Computing*, pages 100–109, April 1983.

[16] Steven Fortune and James Wyllie. Parallelism in random access machines. In *Proceedings ACM Symposium on Theory of Computing*, pages 114–118, 1978.

[17] F. Gavril. Merging with parallel processors. *Communications of the ACM*, 18(10):588–591, 1975.

[18] Hillel Gazit, Gary L. Miller, and Shang-Hua Teng. *Optimal Tree Contraction in the EREW Model*, pages 139–156. Plenum Publishing Corporation, 1988.

[19] L. M. Goldschlager. A universal interconnection pattern for parallel computers. *Journal of the Association of Computer Machinery*, 29(3):1073–1086, 1982.

[20] Allan Gottlieb, R. Grishman, Clyde P. Kruskal, Kevin P. McAuliffe, Larry Rudolph, and Marc Snir. The NYU Ultracomputer—designing a MIMD, shared-memory parallel machine. *IEEE Transactions on Computers*, C-32:175–189, 1983.

[21] Allan Gottlieb, B. D. Lubachevsky, and Larry Rudolph. Basic techniques for the efficient coordination of very large numbers of cooperating sequential processors. *ACM Transactions on Programming Languages and Systems*, 5(2), April 1983.

[22] W. Daniel Hillis. *The Connection Machine*. MIT Press, Cambridge, MA, 1985.

[23] D. S. Hirschberg, A. K. Chandra, and D. V. Sarwate. Computing connected components on parallel computers. *Communications of the ACM*, 22(8):461–464, 1979.

[24] C. A. R. Hoare. Quicksort. *Computer J.*, 5(1):10–15, 1962.

[25] Kenneth E. Iverson. *A Programming Language*. Wiley, New York, 1962.

[26] D.E. Knuth. *Sorting and Searching*. Addison-Wesley, Reading, MA, 1973.

[27] Clyde P. Kruskal, Larry Rudolph, and Marc Snir. The power of parallel prefix. In *Proceedings International Conference on Parallel Processing*, pages 180–185, August 1985.

[28] Richard E. Ladner and Michael J. Fischer. Parallel prefix computation. *Journal of the Association of Computer Machinery*, 27(4):831–838, October 1980.

[29] Frank Thomson Leighton. Tight bounds on the complexity of parallel sorting. In *Proceedings ACM Symposium on Theory of Computing*, pages 71–80, May 1984.

[30] Charles E. Leiserson. Area-efficient layouts (for VLSI). In *Proceedings Symposium on Foundations of Computer Science*, 1980.

[31] Boris D. Lubachevsky and Albert G. Greenberg. Simple, efficient asynchronous parallel prefix algorithms. In *Proceedings International Conference on Parallel Processing*, pages 66–69, August 1987.

[32] G. A. Mago. A network of computers to execute reduction languages. *International Journal of Computer and Information Sciences*, 1979.

[33] Gary L. Miller and John H. Reif. Parallel tree contraction and its application. In *Proceedings Symposium on Foundations of Computer Science*, pages 478–489, October 1985.

[34] William M. Newman and Robert F. Sproull. *Principles of Interactive Computer Graphics*. McGraw-Hill, New York, 1979.

[35] Yu Ofman. On the algorithmic complexity of discrete functions. *Cybernetics and Control Theory, Sov. Phys Dokl.*, 7(7):589–591, January 1963.

[36] G. F. Pfister and V. A. Norton. 'Hot Spot' contention and combining in multistage interconnection networks. In *Proceedings International Conference on Parallel Processing*, pages 790–797, August 1985.

[37] Abhiram G. Ranade. *Fluent Parallel Computation.* PhD thesis, Yale University, Department of Computer Science, New Haven, CT, 1989.

[38] James B. Salem. **Render: A Data Parallel Approach to Polygon Rendering.* Technical Report VZ88–2, Thinking Machines Corporation, January 1988.

[39] Carla Savage and Joseph Ja'Ja'. Fast, efficient parallel algorithms for some graph problems. *Society for Industrial and Applied Mathematics*, 10(4):682–691, 1981.

[40] Jacob T. Schwartz. Ultracomputers. *ACM Transactions on Programming Languages and Systems*, 2(4):484–521, October 1980.

[41] Charles L. Seitz. The Cosmic Cube. *Communications of the ACM*, 28(1):22–33, January 1985.

[42] Y. Shiloach and U. Vishkin. Finding the maximum, merging and sorting in a parallel computation model. *Journal of Algorithms*, 2(1):88–102, 1981.

[43] Yossi Shiloach and Uzi Vishkin. An $O(\log n)$ parallel connectivity algorithm. *Journal of Algorithms*, 3:57–67, 1982.

[44] Harold S. Stone. Parallel processsing with the perfect shuffle. *IEEE Transactions on Computers*, C-20(2):153–161, 1971.

[45] Robert E. Tarjan. *Data Structures and Network Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1983.

[46] Thinking Machines Corporation. Connection Machine parallel instruction set (PARIS). July 1986.

[47] C. D. Thompson and H. T. Kung. Sorting on a mesh-connected parallel computer. *Communications of the ACM*, 20:263–271, 1977.

[48] James C. Wyllie. *The Complexity of Parallel Computations.* Technical Report TR-79-387, Department of Computer Science, Cornell University, Ithaca, NY, August 1979.

Appendix: A Short History of the Scan Operations

This appendix gives a brief history of the scan operations.

Scans were suggested by Iverson in the mid 1950s as operations for the language APL [25]. A parallel circuit to execute the operation was first suggested by Ofman in the early 60s [35] to be used to add binary numbers—the following routine executes addition on two binary numbers with their bits spread across two vectors A and B:

$$(A \otimes B) \otimes \texttt{seg-or-scan}(AB, \overline{A}\,\overline{B})$$

A parallel implementation of scans on a perfect shuffle network was later suggested by Stone [44] to be used for polynomial evaluation—the following routine evaluates a polynomial with a vector of coefficients A and variable x at the head of another vector X:

$$A \times \times\texttt{-scan(copy(}X\texttt{))}$$

Ladner and Fisher first showed an efficient general-purpose circuit for implementing the scan operations [28]. Wyllie first showed how the scan operation can be executed on a linked list using the P-RAM model [48]. Brent and Kung, in the context of binary addition, first showed an efficient VLSI layout for a scan circuit [10]. Schwartz [40] and, independently, Mago [32] first suggested the segmented versions of the scans [40]. More recent work on implementing scan operations in parallel include the work of Fich [15], which demonstrates a more efficient implementation of the scan operations, and of Lubachevsky and Greenberg [31], which demonstrates the implementation of the scan operation on asynchronous machines.

As concerns terminology, *scan* is the original name given to the operation. Ladner and Fisher introduced the term *parallel prefix operation.* Schwartz used the term *all partial sums.* I find the original term, scan, more concise and more flexible—it, for example, can be used as a verb, as in “the algorithm then scans the vector”.

```
define seg-max-scan(A, SFlag){
    Seg-Number ← SFlag + enumerate(SFlag);
    B ← append(Seg-Num, A);
    C ← extract-bot(max-scan(A′));
  if SFlag then C else 0}
```

A	=	[5	1	3	4	3	9	2	6]
SFlag	=	[T	F	T	F	F	F	T	F]
Seg-Number	=	[1	1	2	2	2	2	3	3]
B	=	[1,5	1,1	2,3	2,4	2,3	2,9	3,2	3,6]
C	=	[0	5	5	3	4	4	9	2]
Result	=	[0	5	0	3	4	4	0	2]

Figure 16: The implementation of a segmented `max-scan`.