LABORATORY FOR
COMPUTER SCIENCE

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

MIT/LCS/TM-323

A Space-Efficient Algorithm for Finding the Connected Components of Rectangles in the Plane

Charles E. Leiserson
Cynthia A. Phillips

February 1987

545 TECHNOLOGY SQUARE, CAMBRIDGE, MASSACHUSETTS 02139

A Space-Efficient Algorithm for Finding the Connected Components of Rectangles in the Plane

Charles E. Leiserson
Cynthia A. Phillips

Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

July 1985
(last revised February 1987)

Abstract

We present an algorithm for determining the connectivity of a set of N rectangles in the plane, a problem central to avoiding *aliasing* in VLSI design rule checkers. Previous algorithms for this problem either worked slowly with a small amount of primary memory space, or worked quickly but used more space. Our algorithm uses $O(W)$ primary memory space, where W, the *scan width*, is the maximum number of rectangles to cross any vertical cut. The algorithm runs in $O(N \lg N)$ time and requires no more than $O(N)$ transfers between primary and secondary memory.

Keywords: computational geometry, design rule checking, VLSI, algorithms, rectangles, connected components, scanning.

1 Introduction

For a VLSI design to be reliably produced as a working chip, various features on the chip must be separated by minimum distances to ensure the proper operation of transistors and interconnections. The design rule checker program verifies that these and other geometric constraints are satisfied and signals an error if it finds two features that violate the design rules. For a chip composed of millions of rectangles, design rule checking is a time-consuming process which cannot be done entirely within the primary memory of many computers.

This research was supported in part by the Defense Advanced Research Projects Agency under Contract N00014-80-C-0622.

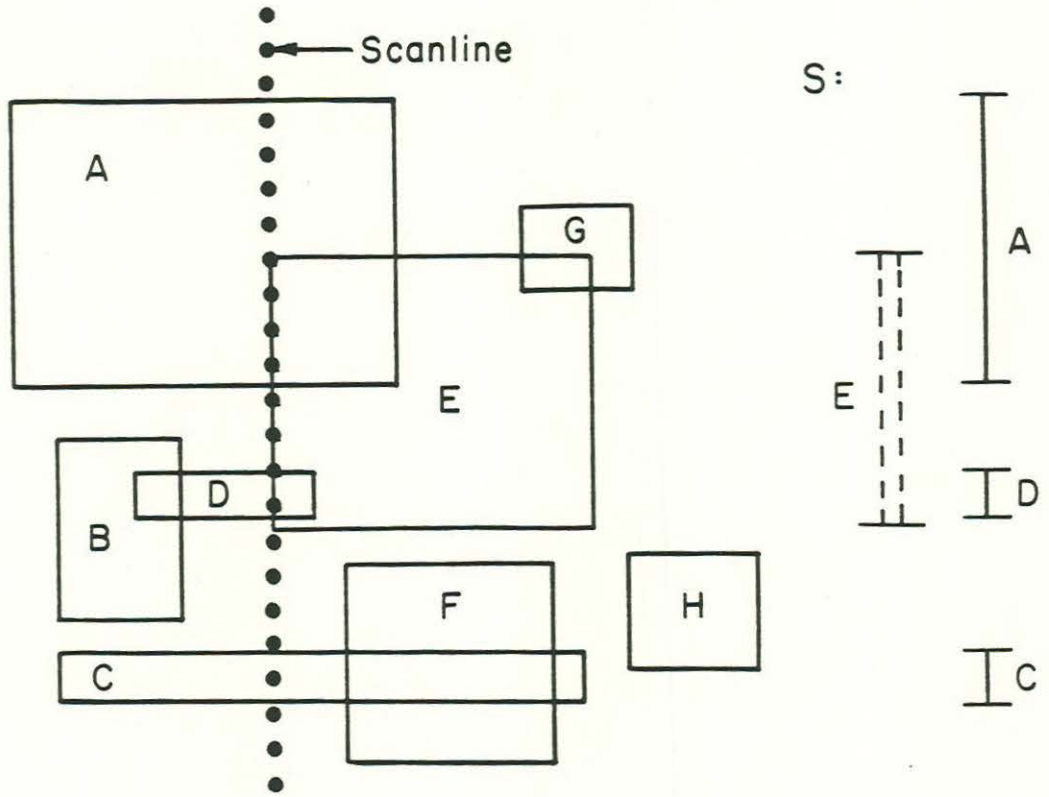

Figure 1: A set of rectangles with connected components $\{A, B, D, E, G\}$, $\{C, F\}$, and $\{H\}$. On the the right is shown a scan set at the time rectangle E enters. Only active rectangles (those crossed by the scanline) have an interval in the scan set. The interval for E will be entered in the scan set S after all processing for its enter event is complete.

This paper presents an efficient algorithm for finding the connected components of rectangles in the plane using a machine model that incorporates the secondary disk memory where the VLSI design is stored. By running this algorithm simultaneously on each layer of a VLSI chip design, a design rule checker can determine which features of a chip design are electrically equivalent, *i.e.*, are effectively part of the same wire. The determination of electrical equivalence allows the design rule checker to avoid reporting the many ***aliasing*** errors that occur when two electrically equivalent features are mistaken for electrically distinct features. For example, two wires might be too close together, but if they are actually the same wire, it does not matter.

Many VLSI design systems use rectilinearly oriented rectangles to represent the design features. Two rectangles are electrically equivalent if they are connected by a path of intersecting rectangles. The connected components problem is to label each rectangle in a design such that two rectangles have the same label if and only if they are in the same connected component. The set of rectangles in Figure 1, for instance, has three connected components: $\{A, B, D, E, G\}$, $\{C, F\}$, and $\{H\}$.

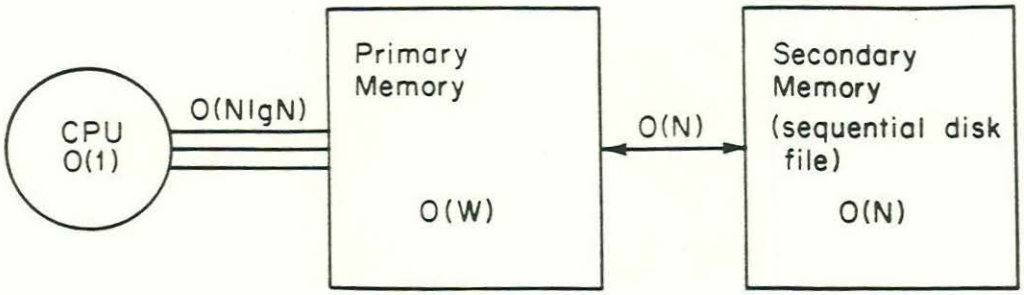

Figure 2: The computer model includes a secondary disk memory as well as primary memory. The connected components algorithm, which assumes the rectangle database is on disk, uses $O(N)$ references to sequential files, $O(W)$ primary memory space, and $O(N \lg N)$ CPU time.

The connected components of N rectangles in the plane can be determined by an algorithm due to Guibas and Saxe [4] in $O(N \lg N)$ time, which is remarkable in that there may be as many as order N^2 rectangle intersections.[1] Their algorithm uses the technique of *scanning,* introduced by Shamos and Hoey [8], which assumes that the vertical edges of rectangles are initially sorted by x-coordinate. Scanning algorithms work by sweeping a *scanline* over a set of geometric objects in the plane and then working primarily with the objects crossed by the scanline. In the Guibas-Saxe algorithm, the scanline is a vertical line that sweeps from left to right over the rectangles. Unfortunately, the Guibas-Saxe algorithm is designed to run entirely within primary memory, and it may cause disk thrashing for a large VLSI chip.

In this paper, we abandon the simple primary memory model, and instead use a machine model that includes a secondary disk memory as well as primary memory. The configuration is shown in Figure 2. We assume that the primary memory is a fast, random-access memory of limited size. The set of rectangles is kept in a file in secondary disk storage. Accesses to the file are presumed to be sequential, either forward or backward. More general random accesses to disk blocks are unnecessary for our algorithm.

This model is used by Szymanski and Van Wyk [9] for a connected components algorithm which is more suitable for large rectangle databases than the Guibas-Saxe algorithm. The Szymanski-Van Wyk algorithm uses less primary memory than the Guibas-Saxe algorithm and has locality of reference for secondary memory. The amount of primary memory space used by the algorithm is $O(W)$, where W, the *scan width*, is the largest number of rectangles cut by any scanline. In practice, Szymanski and Van Wyk comment, the size of W is about $O(\sqrt{N})$. Unfortunately, their algorithm is based on rectangle intersections, and the running time can be as large as $\Theta(NW)$.

This paper presents a connected components algorithm that combines and optimizes

[1] Imai and Asano [5] also have an $O(N \lg N)$ connected components algorithm for the primary memory model which is not based on scanning.

the Szymanski and Van Wyk and the Guibas and Saxe algorithms. It uses $O(W)$ (primary memory) space and runs in $O(N \lg N)$ time in the worst case.

The algorithm consists of a two-pass scan over the set of rectangles. Most of the work is done in the first, *forward* scan. A *backward* scan is then used to produce the labeling of rectangles such that two rectangles have the same label if and only if they are in the same connected component. The algorithm maintains four data structures of size $O(W)$ during its forward scan.

The remainder of this paper presents the connected components algorithm and its analysis. Sections 2.1, 2.2, 2.3, and 2.4 describe the four data structures used during the forward scan. Section 3 gives the algorithm, section 4 proves its correctness, and section 5 analyzes its time and space requirements. Finally Section 6 offers some concluding remarks.

2 Data Structures

In scanning algorithms an *event* is a geometric phenomenon that causes some computation at the time when it occurs. There are two types of events for a left-to-right scan: a *start event* when the scanline crosses the left boundary of a rectangle (the rectangle becomes *active*, or *enters*) and an *end event* when the scanline crosses the right boundary of a rectangle (the rectangle becomes *inactive*, or *leaves*). Each rectangle has an associated start event and end event. The four data structures given in this section are used by the connected components algorithm during scanning.

2.1 The rectangle set

The *rectangle set* R is a dynamic set that contains the active rectangles at any point during the scan. We assume that each rectangle in the disk file has a unique identification number. When a rectangle enters primary memory, it is stored in the set R with the identification number as a key. The rectangle set can be maintained as a balanced search tree, using $O(W)$ space. Each insertion, deletion, or search takes $O(\lg W) = O(\lg N)$ time.

2.2 The scan set

The data structure that maintains the scanline for the connected components algorithm is called the *scan set*. At any point during the forward scan, the active rectangles can be represented as a set of vertical intervals, *i.e.*, an interval in y. For example, Figure 1 shows the intervals of the active rectangles at the time rectangle E enters. The scan set S maintains the dynamic set of intervals that represents the active rectangles.

The scan set allows the connected components algorithm to determine rectangle intersections easily. Two rectangles intersect if and only if there is a scanline that crosses

both rectangles, and their intervals overlap in the scan set corresponding to the scan-line. This technique for determining rectangle intersections is well known and is used in previous scan-based algorithms for determining rectangle intersections or connected components [3,4,9].

To be precise, a scan set S supports the following operations:

S-INSERT!(A): Add rectangle A to the scan set.

S-DELETE!(A): Remove rectangle A from the scan set.

S-FIND(I): Return a rectangle in the scan set S that overlaps interval I in some way, and NIL if no rectangles overlap I.

The number of rectangles stored in S at any given time during a scan is at most the scan width W. We use an *interval tree* [7], a simple, sparse variation on a balanced tree, to implement each of the three operations in time $O(\lg W)$ and space $O(W)$. Alternatively we can achieve the same assymptotic space and time bounds for the above operations by using McCreight's priority search trees [6] to store rectangles keyed on interval.

2.3 Component set

During the forward scan, the connected components algorithm maintains a *component set* Q that reflects our current knowledge of the connectivity of the active rectangles. Each component is designated by a *color,* which for convenience is represented as an integer.[2]

The rectangle colorings within the component set Q may change with a start event. If a new rectangle connects two previously unconnected components, we merge them within the component set Q by recoloring active rectangles in the smaller of the two.

The component set Q supports the following operations:

COLOR!(A): Assigns rectangle A a new (unused) color.

UNCOLOR!(A): Dissociates rectangle A from others of its color. If A is the last of its color, the color is destroyed (made available for reuse).

COLOR(A): Returns A's color.

REPRESENTATIVE(q): Returns any rectangle having color $q \in Q$. If there is no such rectangle, return NIL.

RECOLOR!(q_1, q_2): Takes all rectangles of color q_1 and color q_2 and makes them all either color q_1 or color q_2. The other color is destroyed.

[2]The letter Q is mnemonic for "qonnected qomponents" and "qolor." The first letters of the alphabet are reserved for rectangles.

We implement the component set Q using a vector in which each color is represented as an index in the vector. Each slot in the vector contains a pointer to the first rectangle in a doubly linked list of all rectangles of that color, and the number of rectangles in the list. The pointers to implement the linked lists can be stored with the actual rectangles. Each rectangle also stores the index of its color. If the number field is zero, the color is unused, and we then use the pointer field to implement a free list of the unused colors. An extra variable is needed to store the head of the free list.

All operations except RECOLOR! can be implemented in constant time. If we always merge the color with the smaller number of rectangles into the one with the larger number, then we can do $O(N)$ recolorings in $O(N \lg N)$ time. There are at most W rectangles in the component set Q at any given time so the data structure need only be size $O(W)$.

2.4 Territory set

To achieve an $O(N \lg N)$ worst case running time for the connected components algorithm, we must find a way to maintain the component set Q without looking at every intersection. Figure 3 shows the basic idea. The active rectangles B, C, and D have the same color, say 1. The new rectangle E intersects all three of these rectangles, which tells us that rectangle E should be given the same color as rectangle B, all rectangles with B's color should be merged with rectangles of rectangle C's color, etc. We would get the same result, however, if we just noticed that rectangle E intersects some rectangle(s), all of color 1. That is, instead of asking, "What other rectangles does rectangle E intersect?" we would like to be able to ask, "Is there a color q in the component set Q such that rectangle E intersects at least one rectangle colored q?" We now describe a new data structure called a *territory set* T that allows us to answer this question using small space and time.

The territory set T is a refinement of the *illuminator* data structure used by Guibas and Saxe in their algorithm for the connected components problem [4]. The territory set is essentially a colored partition $\{t_i\}$ of the scanline. Conceptually, each territory has two fields: its *interval* and its *color*. The interval is a closed interval in y. We implement the color indirectly by associating with each territory a *representative rectangle* which is in the territory, and therefore has the same color as the territory. Each territory t in T obeys the following rules:

1. Each active rectangle is covered by exactly one territory.
2. Each territory covers at least one active rectangle. To ensure that the territory set is never empty, we assume there is a dummy rectangle above all rectangles in the data base that extends the full length of the design.
3. All active rectangles covered by territory t have the same color as t.

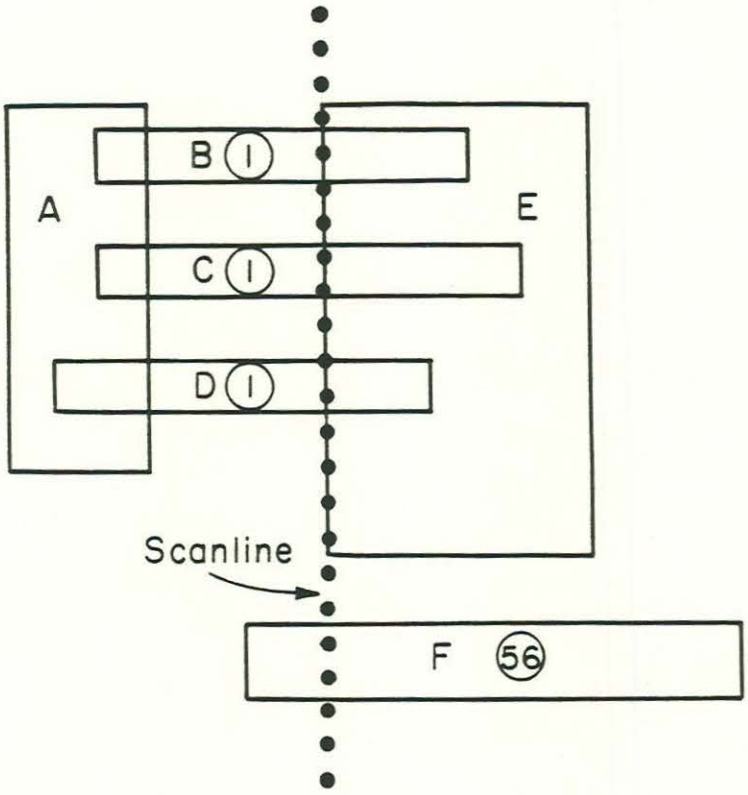

Figure 3: The inefficiencies that can arise from intersection-based connectivity algorithms. Colors of active rectangles are represented as circled numbers. When rectangle E enters we would like to know it should have the same color as each rectangle colored 1 without recognizing this fact three different times via intersection checks.

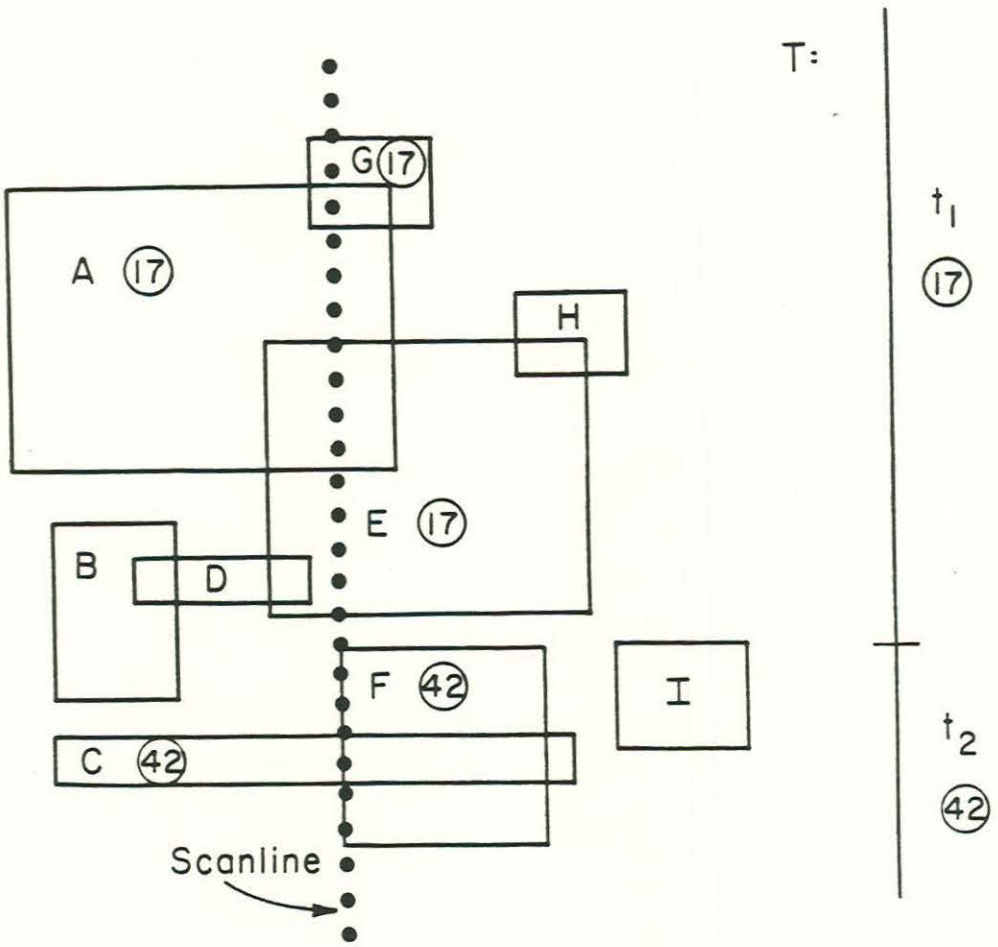

Figure 4: The territory set (right) for a collection of rectangles (left) is essentially a colored partition of the scanline. Colors of active rectangles and territories are represented as circled numbers.

For example, in Figure 4 no rectangles go across the boundary between territories t_1 and t_2. Each territory covers at least one active rectangle. Each active rectangle's color corresponds to the color of the territory that covers it. Here, rectangles A, E and G and territory t_1 that covers them are colored 17. Rectangles C and F and territory t_2 are colored 42.

The territory set T supports the following operations:

T-INSERT!(t): Add territory t to the territory set.

T-DELETE!(t): Delete territory t from the territory set.

LOCATE(y): Return the territory that includes the y-coordinate y. If the point y falls on the boundary between two territories, the lower of the two is returned.

NEXT(t): Return the territory immediately above territory t.

COLOR(t): Return the color of territory t. This operation involves getting t's representative rectangle and getting the color from the rectangle.

The territory set T can be implemented as a standard height-balanced tree using $O(W)$ space. The operations T-INSERT!, T-DELETE!, LOCATE, and NEXT can each be implemented in $O(\lg W) = O(\lg N)$ time. As a simple optimization, the territories can be linked in order, which allows NEXT to run in constant time.

3 The Connected Components Algorithm

This section presents the connected components algorithm which operates in two phases. The first phase is a forward scan over the rectangles during which connectivity information is prepared that is written out to an intermediate sequential file on disk. The second phase is a backward scan over the intermediate file during which component labels are assigned to each rectangle.

The algorithm assumes that the events, which correspond to the scanline crossing left or right edges of rectangles, are sorted by x-coordinate. If not, the events must first be sorted, which takes $O(N \lg N)$ time in the worst case. (As a practical matter, we can often do much better because many VLSI databases already keep rectangles sorted by left edge.) Given a file sorted by left edge alone, we can sort it into start and end events in $O(N \lg N)$ time and $O(W)$ space using an idea due to Szymanski and Van Wyk [9]. The idea is to keep a priority queue, such as a heap [1, pp. 147–152], in primary memory. During the operation of the algorithm, the priority queue holds at most $W + 1$ rectangles sorted by right endpoint. When a new rectangle is read in, its right endpoint is stored in the priority queue. Then the priority queue is emptied of all rectangles with right endpoint smaller than the left endpoint of the new rectangle. For each of these rectangles, the right endpoint is written out in order as an end event. Then the left endpoint of the new rectangle is written out as a start event. Thus, without loss of generality, we can assume the start and end events are presorted.

There are other, more mundane data management issues to be faced in the course of programming the connected components algorithm described here. Most of these can be resolved using simple pointer associations, but the more complicated will be addressed directly in the sections to come.

3.1 The forward scan

The data structures used by the forward scan contain only those rectangles that are active, which ensures that the $O(W)$ space bound is met, but which also leads to problems maintaining connectivity across the entire database. When we see an end event for a rectangle A signaling that A is to become inactive, we are not prepared to give A a final label, yet we must purge A from our internal data structures. For example, at the time

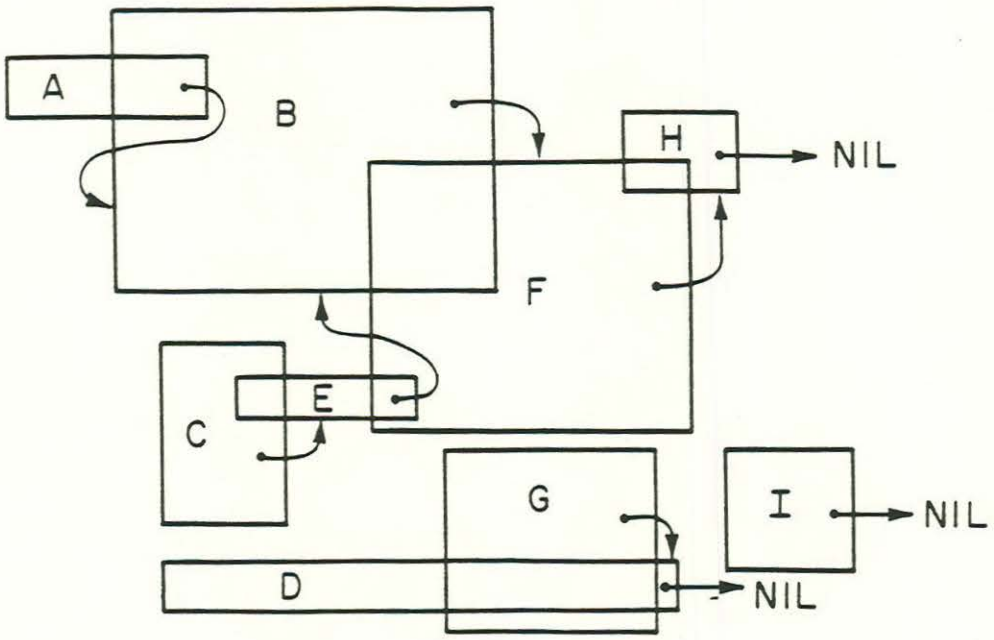

Figure 5: Each rectangle A picks a friend that is active at the time A leaves and is known to be in the same connected component.

rectangles A and C in Figure 5 become inactive, there is no way to guess that they are in the same connected component. Were we to give them final labels now, we would incorrectly give them distinct labels.

Since we cannot give each rectangle A a final label in the forward scan, we give it a *friend*. Rectangle A's friend is another rectangle which (1) is active at the time rectangle A leaves, and (2) is known to be in the same connected component as rectangle A. If there is no such rectangle at the time rectangle A leaves, then its friend is NIL. Figure 5 shows a possible assignment of friends.

At the end of the forward pass, each connected component is linked together by a tree of friend arrows. From this friend information, the back pass can construct final component labels. The idea is that each friend arrow points from left to right if the source and destination rectangles are sorted by right edge, or equivalently, by time of exit. Thus, a component label assigned to the root of the tree will propagate right to left through the tree during the back scan.

The start event

Processing a start event for rectangle A during the forward scan involves four steps: setting up, handling top and bottom boundary conditions, recoloring affected rectangles, and cleaning up.

Set up. Figure 6 shows the important y-coordinates and intervals for the general case. The bottom and top coordinates of A's interval are designated y_{bot} and y_{top}. The

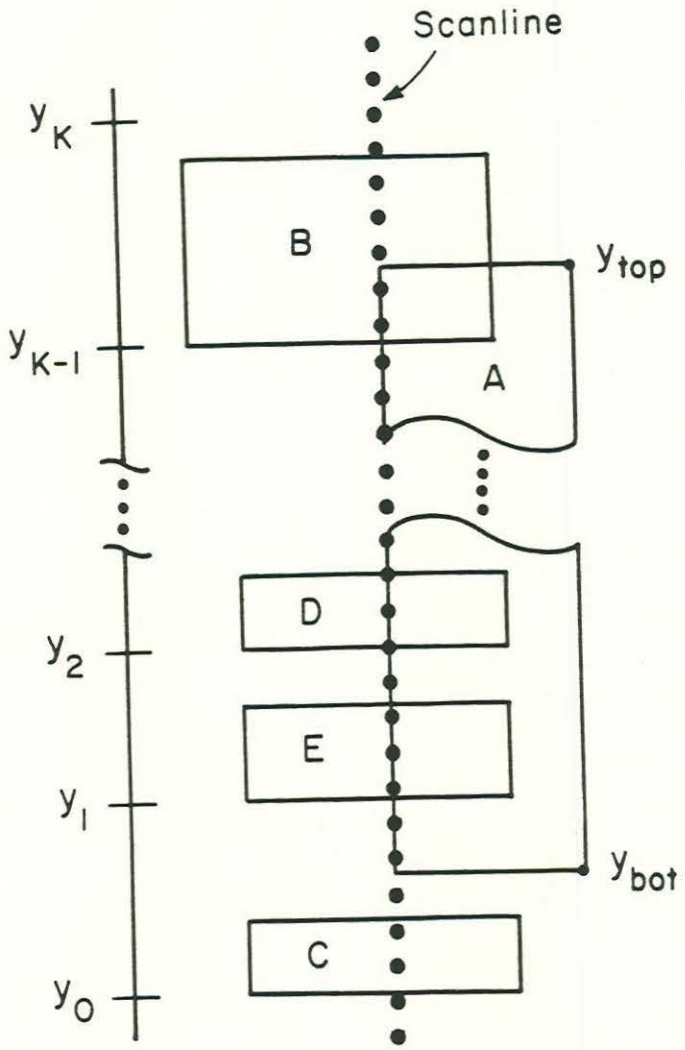

Figure 6: The territory set is shown on the left and the rectangles on the right for the case when $k > 1$. The colors of territories $t_1, t_2, \ldots, t_k$ are a first guess at the colors to merge because of rectangle A's entrance.

endpoints of the k territories in the territory set T that A overlaps are $y_0, y_1, \ldots, y_k$. The k territories are gathered into a list L by first using LOCATEto find the territory that includes y_{bot}, and then using NEXT to gather the remaining territories that overlap A's interval $[y_{bot}, y_{top}]$. All the territories in L are then removed from T, which leaves a gap in T from y_0 to y_k. This gap will be repaired in subsequent steps.

Intuitively, the colors of the territories in list L represent our first guess at which colors must be merged due to the entrance of rectangle A. Since each territory contains at least one active rectangle, the territories in the middle of the list will necessarily contain a rectangle that intersects A.

Handle boundary conditions. Rectangle A extends only partially into the top and bottom territories, so we must explicitly reference the scan set S to determine whether there are active rectangles in these two territories that intersect A. We describe only

the handling of the top boundary condition since the bottom boundary condition is symmetric. Also, for simplicity, we shall consider the special case $k = 1$ (Figure 7) after we deal with the general case $k \geq 2$ (Figure 6).

Handling the top boundary condition for $k \geq 2$ involves determining whether the top territory should be kept in list L. The first case is when there is some active rectangle B that intersects the interval $[y_{k-1}, y_{top}]$. The interval of the rectangle B falls entirely within the top territory, so it follows that A, B, and every other active rectangle covered by the top territory must have the same color by the time we finish processing the entrance of A. Therefore, we leave the top territory in the list L, and nothing is to be done.

Otherwise, no active rectangle intersects A in the top territory, and the top territory is removed from L. Since k is at least 2, there must be an active rectangle in the interval $[y_{top}, y_k]$ because the top territory must contain at least one rectangle, and the interval $[y_{k-1}, y_{top}]$ contains none. Therefore, we can return the top territory to the territory set with the shortened interval $[y_{top}, y_k]$ without violating any of the properties a territory must have. In other words, chopping off empty space does not hurt.

We now discuss the processing of the top boundary condition for the special case when $k = 1$ (Figure 7), since once again, the bottom boundary condition is symmetric. If rectangle A intersects some active rectangle, then we are done. If rectangle A does not intersect any active rectangle, it is possible that the rectangle that justified the existence of the single territory in list L is below rectangle A, instead of above. In this case, we must explicitly query the scan set S with the interval $[y_{top}, y_1]$ to determine whether there is an active rectangle to justify putting a territory over the interval. If there is an active rectangle, we must enter a new territory into T with the shortened interval $[y_{top}, y_1]$ using the color of the old territory.

Recolor. Now, the colors of the territories in list L are exactly the colors that must be merged because of rectangle A's entrance. We first color rectangle A with a new color. We then merge A's color with the color of each territory in L. The colors are automatically garbage collected by the component set Q. Because of our pointer implementation of territory colors, no territory is ever colored with a garbage-collected color.

Clean up. We finish the servicing of rectangle A's entrance by repairing the territory set T and making A active. The gap left after handling boundary conditions becomes the interval of a new territory with the color of rectangle A. We insert Rectangle A into the rectangle set R and the scan set S. Since the left side of a rectangle indicates an end event in the back scan, we enter an end event for rectangle A in the intermediate file that will serve as input to the back scan.

The end event

Servicing an end event for rectangle A requires us first to find the associated rectangle object for A in the rectangle set R. Then, we must output a start event for A in the back pass and fix up the internal data structures. We accomplish this processing in three

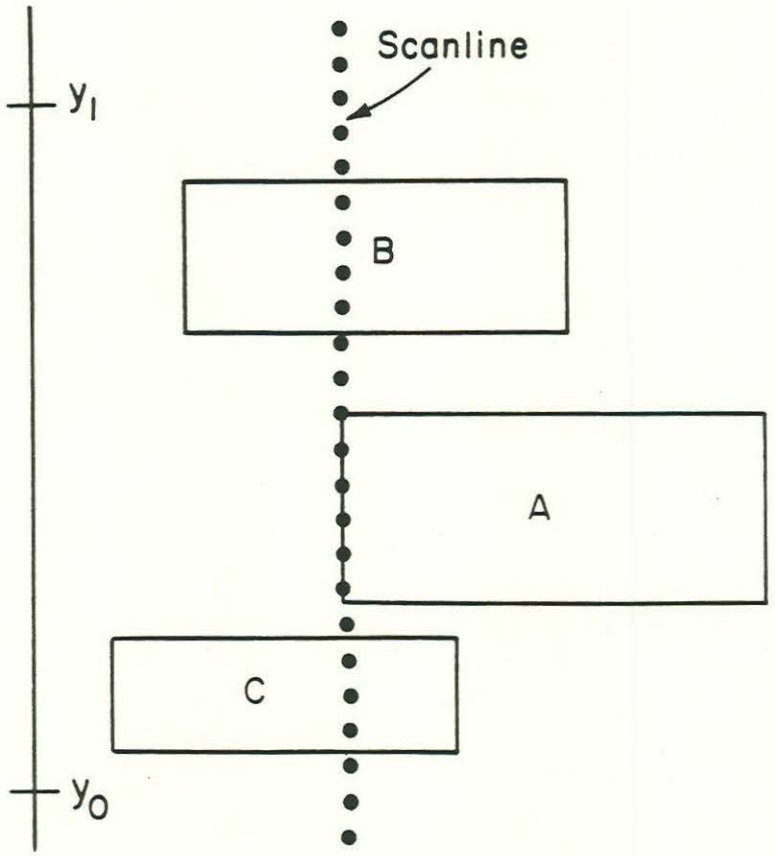

Figure 7: The territory set for the case when $k = 1$. Since rectangle A falls in only one territory, the intervals above and below rectangle A must be checked explicitly.

steps: making rectangle A inactive, associating A with a friend, and fixing the territory set T.

Make A inactive. Let q be rectangle A's color before processing the end event. We uncolor rectangle A, and remove it from the scan set S and the rectangle set R.

Find a friend. We query the component set Q for a representative of color q and associate this representative rectangle (possibly NIL) with rectangle A so that A can now tell its friend when asked. We write out this information as a start event for rectangle A for use in the back scan. We shall say that rectangles that receive NIL as a friend have no friend or are friendless.

Fix the territory set. We pick any point on rectangle A's interval, and use LOCATEto find the one territory t that covers rectangle A. We then find a rectangle B in the scan set S that intersects t's interval to see if there is some active rectangle to justify t's existence. (Recall that a territory must cover at least one active rectangle.) If no rectangle exists, then A is the last active rectangle in t's interval, and territory t can be eliminated by extending the interval of the next territory above t to include t's interval.

If the existence of territory t is justified by some active rectangle B, and A is serving as the representative for territory t, then we make rectangle B the representative of territory t.

3.2 The back scan

The second phase, the back scan, passes backwards through the intermediate file of rectangle-friend information created in the forward scan, and produces a final file of rectangle-label pairs which will be sorted by left edge. During this right-to-left scan, each rectangle receives its final labeling from its friend. The back scan uses only one data structure, the rectangle set R. It also requires a counter initialized to 0.

During the back scan, the rectangle set R holds all active rectangles, and each active rectangle knows its final component label. Labels are assigned sequentially during the back scan, and the counter holds the value of the next label to be assigned.

The start event

The first step in servicing a start event for rectangle A is to assign a final label to A. If A has no friend, it is the rightmost rectangle in its component, and so a new label must be assigned from the counter. We store this label into A and increment the counter.

Otherwise, find rectangle A's friend in the rectangle set R, and give A the same label as its friend. Rectangle A's friend must be active since A and its friend were simultaneously active in the forward scan. Rectangle A left first in the forward scan so it must enter after its friend in the back scan. Finally, we add rectangle A to the rectangle set R.

The end event

Processing an end event for a rectangle A consists of simply removing A from the rectangle set R and writing out rectangle A with its label to a final file. No rectangle that subsequently enters has A as a friend because the two rectangles are not simultaneously active. Thus, no other rectangle will need to get a label from A, and hence, it is safe to remove A. The final file is sorted by left edge from right to left. Reversing the file leaves it sorted left to right by left edge as was the original input file.

4 Proof of correctness

This section shows that two rectangles get the same label if and only if they are in the same connected component.

($\Rightarrow$) We first show that if two rectangles are given the same label, then they are in the same connected component. We prove this by induction on the number of rectangles given the same label. Suppose rectangle A is the first rectangle given label l. Then at the time we process rectangle A's start event during the back scan, rectangle A is friendless and the value of the counter is l. If rectangle A had a friend, it would be given the same label as its friend contradicting our assumption that rectangle A was the first to receive its label. The counter is incremented after rectangle A is given the label l so no friendless rectangles to enter after A get the label l. By the same argument, no friendless rectangles to enter before rectangle A are given the label l.

Assume that at some point in the backscan, k rectangles have been given label l and all k are in the same connected component. Some $j \leq k$ of these rectangles are active (*i. e.* are in the rectangle set R). Now the start event for some rectangle B causes B to get label l. For this to happen, rectangle B must have a friend rectangle C which is one of the j active rectangles with label l. Since rectangle C is rectangle B's friend, both rectangles B and C must have had the same color in the component set Q at the time rectangle B left in the forward scan.

To finish the argument, we show that two rectangles simultaneously having the same color in the component set Q are in the same connected component. If this is true, then rectangles B and C are in the same connected component and therefore by transitivity rectangle B is in the same connected component as the other k rectangles given label l.

We show that two rectangles sharing a color in the component set Q must be in the same connected component by induction on the number of rectangles with that color. A new color is introduced into the component set Q only when a COLOR! operation is performed upon a rectangle A during its start event. Hence each color begins with only one member rectangle. Other rectangles join a color only through the RECOLOR! operations performed during the processing of the start event for a rectangle.

Assume that before processing the start event for a rectangle A, all rectangles with the same color in the component set Q are in the same connected component. After

handling the boundary conditions, there are $m \geq 0$ territories in the list L. These territories have $n \leq m$ distinct colors $q_1, q_2, \ldots, q_n$ which are all merged into one final color q. The colors $q_1, q_2, \ldots q_n$ are exactly those colors for which at least one member rectangle intersects rectangle A. Each pair of rectangles in the final color q is connected by a path of intersecting rectangles. If they shared a color q_i in the component set Q before rectangle A entered, then by assumption there is a path connecting them that includes only rectangles originally colored q_i. Otherwise, there is a path between the two rectangles that includes rectangle A. Therefore all rectangles now colored q are in the same connected component.

($\Leftarrow$) We now prove that if two rectangles are in the same connected component, then they get the same label. It suffices to show that if two rectangles intersect, they get the same label because then all rectangles in the same connected component get the same label by transitivity. The proof has two parts. First, we argue that if two rectangles intersect, then during the forward scan they have the same color in the component set Q while they are both active. Next we show that if two rectangles are simultaneously active and have the same color in the component set Q, then they get the same label.

To show that if two rectangles A and B intersect, they have the same color in the component set Q while they are both active, assume without loss of generality that rectangle B enters after rectangle A. Let t be the territory in the territory set T that covers rectangle A at the time rectangle B enters. Since rectangles A and B intersect, territory t must at least partially cover rectangle B so it is gathered into the list L in the first step of the processing of the start event for rectangle B. The presence of the active rectangle A intersecting rectangle B guarantees that after the boundary condition checks, territory t is still in the list L. Therefore after the merging, rectangles A and B are the same color in the component set Q. From that point on they always move together in any recolorings, so they always have the same color until one of them leaves.

To show that if two rectangles are in the same color in the component set Q while they are active, then they get the same final label, suppose a rectangle A is about to leave, and consider the set of rectangles that have the same color as rectangle A. Rectangle A chooses one as a friend (shown by an arrow in Figure 8). Later, other rectangles may join this set through merges. As each rectangle in the set leaves, it chooses a friend from among those left in the set. Eventually an exiting rectangle finds itself alone, and it exits without a friend. For example, Figure 8 illustrates the sequence of friend choices for one set of rectangles taken from the example in Figure 5. During the forward scan each of these rectangles simultaneously shares a color in the component set Q with at least one other rectangle in the set. For example rectangles A and B share a component immediately after rectangle B enters and rectangles B, E, and F share a component immediately after rectangle F enters.

We can view the illustration in Figure 8 as an acyclic graph with the rectangles as vertices and the friend relation arrows as directed edges. Each vertex has outdegree one except for a single sink, the friendless rectangle H. If we start at any vertex in the graph and follow the edges, we always end up at the sink. We know from our previous argument that a rectangle gets the same label as its friend. That friend in turn gets

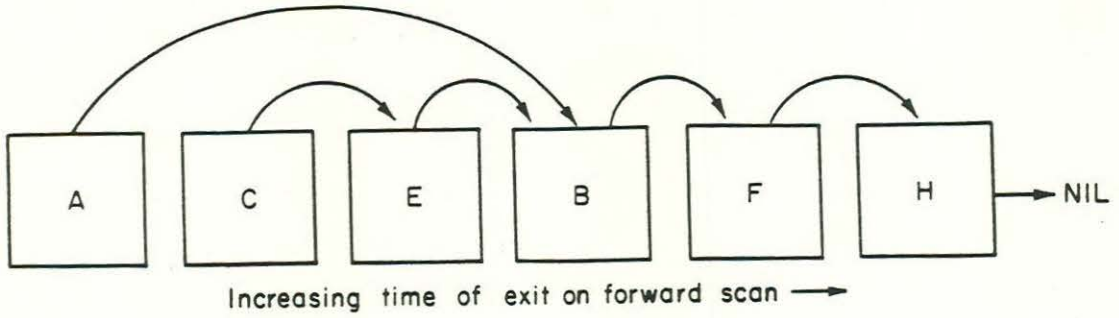

Figure 8: An example of a component taken from Figure 5. The arrows represent the friend relation. Each rectangle shares a component color with at least one other rectangle in this set during the forward scan. During the back scan, these rectangles enter from right to left. Rectangle H receives a new label, and all the other rectangles receive their labels indirectly from rectangle H.

the same label as *its* friend, ... (down the friend links) ..., who gets the same label as the sink H. By transitivity any two rectangles that are in the same component of the component set Q while active get the same final label.

5 Analysis

This section shows that the worst-case running time of the connected components algorithm is $O(N \lg N)$, the amount of primary memory required is $O(W)$, and the number of transfers between primary and secondary memory is $O(N)$. We have already seen that each data structure requires only $O(W)$ primary memory space, and it can be verified that the number of disk transfers is $O(N)$. Thus, we must demonstrate that the running time of the algorithm is $O(N \lg N)$.

The rectangle set R and the scan set S each contribute only $O(N \lg N)$ to the overall time. The rectangle set R is used in both the forward scan and the back scan. It contributes only $O(N \lg N)$ to the time in each phase since it performs at most two operations, each requiring $O(\lg N)$ time, on each of the $O(N)$ start and end events. The scan set S performs one insertion or deletion and at most four S-FIND operations for each start or end event.

Operations on the territory set T contribute $O(N \lg N)$ time as well. During the servicing of an end event, the territory set T performs at most one LOCATE, two T-DELETE!'s, and one T-INSERT!, if we regard the modification of a territory interval as a deletion followed by an insertion. For a start event, only one LOCATE and at most three T-INSERT!'s are performed. The number of calls to NEXT, T-DELETE!, and COLOR directly depends on the size of the list L, however.

We shall show that each operation is performed at most $O(N)$ times. The operations

that are performed a constant number of times for a given event are executed $O(N)$ times overall. The other operations are called once for each time a territory appears in a list L during a start event. Thus, showing that the sum total of the sizes of L throughout the entire forward pass is $O(N)$ will produce our desired bound. The total number of insertions into T is at most $4N$, which therefore bounds the total number of deletions. Moreover, each of these territories can participate in a list L only once since it is deleted from T at that time and replaced by the consolidated territory or a new boundary territory. Hence, the sum total of the lengths of L is $O(N)$, which also bounds the number of times any operation is performed. Since each operation costs $O(\lg N)$ time, the total work performed on the territory set is $O(N \lg N)$.

It remains to analyze the component set Q. Each start event causes one COLOR! operation, and each end event causes one UNCOLOR! and REPRESENTATIVE operation. Using the same arguments as above for the territory set, at most $O(N)$ RECOLOR! and COLOR operations are performed throughout the whole forward scan. Thus, its contribution to the overall running time of the connected components algorithm is also $O(N \lg N)$.

6 Conclusions

This section presents the important extension of the connected components algorithm to multiple layers. We also discuss some alternative implementations of the data structures which may be better suited to a practical implementation.

The connected components problem of rectangles in the plane presented in this paper is a simplification of the problem faced in computer-aided design of VLSI. Computing the electrically equivalent rectangles in multiple planar layers of a VLSI design is not much more difficult than the one-layer problem discussed in this paper, even though contact cuts can allow components to snake up and down among layers.

To find the connected components of rectangles on multiple layers, we simply run a copy of the basic, one-layer, connected components algorithm on each layer. In the forward scan, each layer is given its own scan set, rectangle set, and territory set. The component set, however, is global to the entire computation. Each contact is represented explicitly on the layers it intersects. In the back scan, both the counter and the rectangle set are global. No further changes are necessary.

Some of the data structures necessary for the connected components algorithm can be implemented more practically than with the asymptotically efficient height-balanced trees presented in the body of the paper. The rectangle set R, for example, can be implemented by hashing on the rectangle identification number, which would lead to good average case behavior. At the cost of a bit more complication, the component set Q can be implemented with a union-find structure that allows $O(N)$ merges in almost linear time [10]. These modifications improve the expected-time performance of some of the bookkeeping operations independent of the statistical distribution of rectangles.

With some assumptions about the distribution of rectangles, the scan set S and the territory set T can also be implemented more efficiently using bins, as has been done for other VLSI algorithms [2]. Each bin represents a fixed portion of the scanline and contains a pointer to the list of objects that overlap that bin. A desirable bin size can be chosen based on statistical information about the VLSI design. The worst-case performance of the algorithm may be diminished, however, because long, tall rectangles are split across many bins, and consequently, the practical difference between this binning approach and Szymanski and Van Wyk's approach [9] may be negligible.

References

[1] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison–Wesley Publishing Co., Reading, MA, 1974.

[2] J. Bentley, D. Haken, and R. Hon, "Statistics on VLSI Design," *Carnegie-Mellon University Department of Computer Science Technical Report No. CMU–CS–80–111*, April, 1980.

[3] J. Bentley and D. Wood, "An optimal worst case algorithm for reporting intersections of rectangles," *IEEE Transactions on Computers*, Vol. C–29, No. 7, July, 1980.

[4] L. Guibas and J. Saxe, "Problem 80–15," *Journal of Algorithms,* Vol. 4, 1983, pp. 177–181.

[5] H. Imai and T. Asano, "Finding the connected components and a maximum clique of an intersection graph of rectangles in the plane," *Journal of Algorithms,* Vol. 4, 1983, pp. 310–323.

[6] E. McCreight, "Priority search trees," *Xerox Corporation Palo Alto Research Centers Technical Report,* CSL–81–5, January 1982.

[7] C. A. Phillips, "Space-efficient algorithms for computational geometry," Master's Thesis, MIT VLSI memo No. 5-270, 1985.

[8] M. I. Shamos and D. Hoey, "Geometric intersection problems," *Proceedings of the Seventeenth Annual Symposium on Foundations of Computer Science,* IEEE, 1976.

[9] T. G. Szymanski and C. J. Van Wyk, "Space efficient algorithms for VLSI artwork analysis," *Proceedings of the Twentieth Design Automation Conference,* June 1983, pp. 734–739.

[10] R. Tarjan, "Efficiency of a good but not linear set union algorithm," *Journal of the Association for Computing Machinery,* Vol. 25, No. 2, 1975, pp. 215–225.