may be used. In the limit, the complexity of the estimation procedure reaches that of a fault simulator. However, it appears that we can obtain reasonable estimates and yet restrict the complexity close to linear in the number of lines since the segment length does not have to be increased with the circuit depth. The segment length can be reduced by using an efficient graph representation of the circuit where nonfan-out nodes have been collapsed. Such a graph preserves the path structure [6]. An interesting problem is to find the optimum segment length for a given circuit structure such that the estimation error is minimized or eliminated.

Test set reordering [1] and polynomial time algorithms to find maximum cardinality cutsets [5] are other alternatives to improve the accuracy of count-based estimators. However, our method of improving the accuracy of the estimation procedure is less susceptible to the nature of the circuit and the test set used, and hence, may be a viable alternative.

References

[1] I. Pomeranz and S. M. Reddy, "An efficient nonenumerative method to estimate the path delay fault coverage in combinational circuits," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 240–250, Feb. 1994.

[2] K. Heragu, "Approximate and statistical methods to compute delay fault coverage," M.S. thesis, Dept. ECE, Rutgers Univ., Piscataway, NJ, May 1994.

[3] K. Heragu, M. Bushnell, and V. D. Agrawal, "An efficient path delay fault coverage estimator," in *Proc. 31st Design Automation Conf.*, June 1994, pp. 516–521.

[4] K. Heragu, J. H. Patel, and V. D. Agrawal, "Improving accuracy in path delay fault coverage estimation," in *Proc. 9th Int. Conf. VLSI Design*, Jan. 1996, pp. 422–425.

[5] D. Kagaris, S. Tragoudas, and D. Karayiannis, "Improved nonenumerative path-delay fault-coverage estimation based on optimal polynomial-time algorithms," *IEEE Trans. Computer-Aided Design*, vol. 16, pp. 309–315, Mar. 1997.

[6] M. Gharaybeh, M. Bushnell, and V. D. Agrawal, "An exact non-enumerative fault simulator for path-delay faults," in *Proc. Int. Test Conf.*, Oct. 1996, pp. 276–285.

The Inversion Algorithm for Digital Simulation

Peter M. Maurer

***Abstract*—The inversion algorithm is an *event-driven* algorithm, whose performance rivals or exceeds that of levelized compiled code simulation, even at activity rates of 50% or more. The inversion algorithm has several unique features, the most remarkable of which is the size of the run-time code. The basic algorithm can be implemented using no more than a page of run-time code, although in practice, it is more efficient to provide several different variations of the basic algorithm. The run-time code is independent of the circuit under test, so the algorithm can be implemented either as a compiled code or an interpreted simulator with little variation in performance. Because of the small size of the run-time code, the run-time portions of the inversion algorithm can be implemented in assembly language for peak efficiency, and still can be retargeted for new platforms with little effort.**

I. Introduction

Of all the tools available to the modern very large scale integration (VLSI) designer, simulation is probably most important. The cost of fabricating a VLSI design is so high that it is necessary to verify and debug the product before committing it to silicon. Despite steady improvements in simulator performance, it is not unusual for a VLSI designer to spend more time on simulation than on any other activity. There are many different styles of simulation, from high-level simulation at the algorithmic level, to electrical simulation using systems of differential equations. As a general rule, the more detailed the simulation, the more time consuming it becomes. Logic simulation represents a compromise between the extremes of algorithmic simulation and electrical simulation. Although more time consuming than algorithmic simulation, it is efficient enough to be used as a primary debugging tool. While it is not as detailed as electrical simulation, the logic gates that comprise the logic model can be mapped one-to-one into the electrical components of the final product.

Over the past several years, there has been a steady flow of papers describing new more efficient methods of logic simulation [1]–[14]. This research makes it obvious that there are two methods for improving the performance of logic simulation: speed up the simulation of individual gates, or simulate fewer gates. Until now, these two methods have worked at cross purposes to each other. Some simulators have used relatively complex algorithms for reducing the number of gates simulated, thereby increasing the simulation time for each gate. Other simulators have improved the speed of individual gate simulations by reducing or eliminating scheduling code, thereby increasing the number of gate simulations that must be performed for each input vector. When all scheduling code is eliminated, the simulation time for each input vector becomes constant, and is no longer dependent on changes in the inputs. Such simulators are termed *oblivious*. In contrast, simulators whose performance varies from one input vector to another are termed *event-driven* [2]. (This term is used for simplicity and does not necessarily imply that the simulator processes events.)

Manuscript received June 7, 1995; revised June 20, 1996 and July 7, 1997. This work was supported in part by the National Science Foundation under Grant MIP-9403414 and by the University of South Florida Center for Microelectronics Research. This paper was recommended by Associate Editor K. Mayaram.

The author is with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620 USA.

Publisher Item Identifier S 0278-0070(97)07566-0.

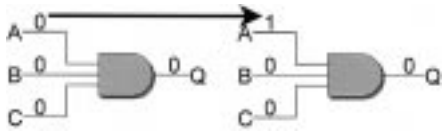


Fig. 1. Useless simulation.

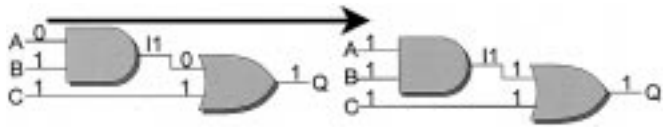


Fig. 2. Unpropagated change.

The simplest form of oblivious simulation is levelized compiled code (LCC) simulation, in which each gate in the circuit is simulated once per input vector. Many of these gate simulations produce no useful output because they do not cause a change in any output net. (Any net visible to the user can be considered an output net, regardless of whether it is an actual output of the circuit.) Event-driven simulation is the most common method for eliminating useless simulations. In a typical event-driven simulation, an event is generated whenever a net changes value. A gate is simulated if and only if an event occurs on one of its inputs. Although this technique eliminates many useless simulations, it does not eliminate all of them. Even if the inputs of a gate change, this change may not propagate to an output net. This can occur in two ways. First, the change in the gate input may not produce a change in the gate output. This situation is illustrated in Fig. 1. Second, the change may propagate through the gate, but be "absorbed" by some other gate before reaching an output net, as illustrated in Fig. 2.

The *inversion algorithm* is able to eliminate all useless simulations of the first kind, and some useless simulations of the second kind. The problem of eliminating *all* useless simulations is currently under study.

II. An Overview of the Inversion Algorithm

Because the aim of the inversion algorithm is to eliminate gate simulations that produce no change in the output of a gate, it was designed around the underlying principle that no gate should be simulated unless its output is guaranteed to change value. Thus, when an event occurs on the input of a gate, it is necessary to determine whether the change will propagate through the gate, and suppress the gate simulation if no propagation will occur. It is not immediately clear that making such a determination is any more efficient than simply simulating the gate, but it turns out that this is indeed the case.

When an event occurs on a gate input, the inversion algorithm performs a set of tests to determine if the event will propagate through the gate. The tests must be individualized for different gate types, but the number of different kinds of tests that must be performed is surprisingly small. In its most basic form, the inversion algorithm supports the eight gate types AND, NAND, OR, NOR, XOR, XNOR, NOT, and BUFFER. These gate types provide all essential functions, and can be used as building blocks to construct more complex gate types. However, it is not necessary to provide specific tests for each of these eight types. One set of tests is provided for NOT and BUFFER gates, a second set of tests is provided for XOR and XNOR gates, and a third set of tests is provided for AND, NAND, OR, and NOR gates.

The tests for the XOR, XNOR, NOT, and BUFFER gates are trivial

```
NOT/BUFFER:
  Schedule OutputEvent;

XOR/XNOR:
  if OutputEvent Not Scheduled Then
      Schedule OutputEvent;
  else
      Deschedule OutputEvent;
  Endif
```

Fig. 3. Event handlers for NOT/XOR.

```
If Value.of.X = Dominant.Value.of.G Then
    Count.of.G := Count.of.G + 1;
    If Count.of.G = 1 Then
        Output.of.G := Dominant.Value.of.G;
    EndIf;
Else
    Count.of.G := Count.of.G - 1;
    If Count.of.G = 0 Then
        Output.of.G := NOT Dominant.Value.of.G;
    EndIf;
EndIf;
```

Fig. 4. Original counting algorithm.

because any change in an input implies a change in the output. An identical set of tests could be used for all four gate types, but because XOR and XNOR gates have more than one input, it is possible to propagate two or more simultaneous events through the gate. Because two consecutive simultaneous events cancel one another, the tests for XOR and XNOR have been optimized to eliminate consecutive events on the gate output. When an event propagates through the gate, the algorithm tests the queue to determine whether there is already an event queued for the net. If so, the existing event is removed from the queue, and no new event is queued. Since NOT and BUFFER gates have a single input, no test for collapsed events is necessary. The event handlers for NOT/BUFFER gates and XOR/XNOR gates are illustrated in Fig. 3.

The tests for AND, OR, NAND, and NOR are more complex, and are based on the counting algorithm originally described by Schuler [16]–[18]. The counting algorithm, which is illustrated in Fig. 4, was originally intended for use in a conventional event-driven simulation. In Fig. 4, it is assumed that there has been a change in an input X to the gate G. The dominant value, 1 for OR/NOR and 0 for AND/NAND, is a parameter to the algorithm.

The counting algorithm of Fig. 4 assigns a value to the output of G if and only if the output changes value. This algorithm is extremely efficient because it uses the value of a single input and an internal state to compute the value of the output, rather than computing a function using all input values. The counting algorithm used by the inversion algorithm is used to predict changes rather than compute output values, and is much simpler than the algorithm shown in Fig. 4. One reason for the simplification is the underlying scheduling technique, which is based on the shadow algorithm [12]. In the shadow algorithm, each event is represented by the structure pictured in Fig. 5.

The inversion algorithm generates a dedicated event structure for each input net, each of which contains an indirect pointer to the event-processing routine for the net. The use of indirect pointers allows event handlers to be changed at run time. In the counting algorithm, the test for dominant values can be eliminated by observing that successive events on an input net will cause it to alternate between its dominant and nondominant value. The inversion algorithm uses two event handlers, one for dominant values and one for nondominant

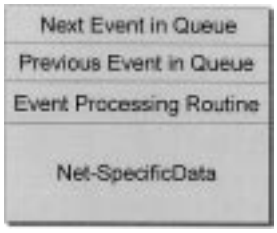


Fig. 5. Event structure.

```
Dominant:
Decrement DominantCount;
If DominantCount=0 Then
    If OutputEvent Not Scheduled Then
        Schedule OutputEvent
    Else
        Deschedule OutputEvent
    EndIf
EndIf
EventProcessingRoutine := AddressOf NonDominant;

NonDominant:
Increment DominantCount;
If DominantCount=1 Then
    If OutputEvent Not Scheduled Then
        Schedule OutputEvent
    Else
        Deschedule OutputEvent
    EndIf
EndIf
EventProcessingRoutine := AddressOf Dominant;
```

Fig. 6. AND/OR event processors.

values. When the dominant-value event-handler executes, it replaces the event-processing-routine address in the event structure with the address of the nondominant event handler. The nondominant-value event handler performs similarly. The two event handlers execute in strict alternating fashion for each input net, with no test for dominant value required. The event processing for AND, NAND, OR, and NOR gates uses the same event-collapsing procedure as XOR and XNOR gates. The event handlers for dominant and nondominant values are illustrated in Fig. 6.

The event handlers of Figs. 3 and 6 do not contain separate code for scheduling gate simulations. Because no gate is scheduled for simulation unless its output is guaranteed to change value, gate simulations are reduced to simple inversion operations. (This assumes that a two-valued logic model is being used. The inversion algorithm supports more complex logic models, but this is beyond the scope of this paper [19].) Surprisingly, because the correct operation of the inversion algorithm does not require net values, it is possible to eliminate most gate simulations entirely. The event handlers pictured in Figs. 3 and 6 do not need to test net values to schedule new events. There are only two cases where net values are required by the inversion algorithm. Net values are required for all primary inputs because it is necessary to compare new and old net values when a new input vector is read. It is also necessary to maintain net values for any net visible to the user so that correct output values can be printed after an input vector has been simulated. Since the processing of events does not depend on net values, gate simulations can be performed during event processing without affecting the correctness of the algorithm. Thus, when the output of a gate is visible to the user, the event handlers will schedule a special event to invert the value of the output.

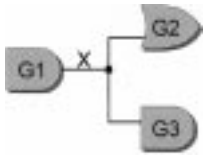


Fig. 7. Circuit fragment.

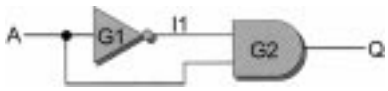


Fig. 8. Complex initialization requirements.

The elimination of net values has some surprising consequences which are worth noting. Inverted outputs and noninverted outputs are identical. Therefore, for simulation purposes, AND is identical to NAND, OR is identical to NOR, XOR is identical to XNOR, and NOT is identical to BUFFER. Because the increment and decrement operations are performed with respect to dominance rather than specific values, AND and OR gates are also identical for simulation purposes.

In a sense, the inversion algorithm performs *no* traditional gate evaluations, but simply processes a series of events. These events differ in one important way from the events that occur in traditional event-driven algorithms. In traditional event-driven simulation, each event corresponds to a change in a single net, while in the inversion algorithm, each event corresponds to a change in a single fan-out branch of a net. Thus, a single event in a traditional event-driven algorithm may correspond to several events in the inversion algorithm. Fig. 7 illustrates why this is necessary.

In Fig. 7, $\boldsymbol{X}$ is the output of gate $G1$, and the input for $G2$ and $G3$. However, the dominant value for $G2$ is the nondominant value for $G3$, and vice versa, so when an increment operation is performed for $G2$, a decrement operation must be performed for $G3$. Although both of these operations could be performed during the processing of a single event, it is more straightforward to treat them as separate events. This implementation style also facilitates the incorporation of inversion events for computing required net values. In Fig. 7, if net X were visible to the user, the simulator would add a third event to compute the value of X.

Initialization for the inversion algorithm is somewhat more complex than for more conventional simulation algorithms. Although the dominant, nondominant sequence is predictable for the inputs of AND, NAND, OR, and NOR gates, it is necessary to commence the simulation with the correct event handler. As Fig. 8 illustrates, a simple default is not sufficient to guarantee correct operation of the algorithm.

In Fig. 8, any time net A changes to the dominant value, net $I1$ changes to the nondominant value, and vice versa. It is necessary to initialize the simulation in such a way as to guarantee that any time the nondominant event handler is used for net A, the dominant event handler will be used for net $I1$. The determination of this requirement cannot be made simply by examining gate $G2$. It is necessary to examine the entire circuit to determine the correct initialization state for the inputs of $G2$. To determine the correct initialization for all gate inputs, a single simulation is performed at compile time to determine consistent values for all nets in the circuit. These net values are then used to determine the initial event handler to be used for each fan-out branch and the initial dominant count for each gate. (The three-valued inversion algorithm eliminates the preliminary simulation step, but this is beyond the scope of this paper.)

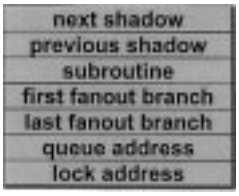


Fig. 9. Structure of a shadow.

III. Implementation Details

The inversion algorithm consists of two major phases, the translation phase which prepares the circuit for simulation, and the simulation phase which performs the simulation. The primary function of the translation phase is to prepare the data structures used by the simulation phase. Most current implementations of the inversion algorithm also generate run-time code, but this code could just as easily be loaded from a library of precompiled routines.

The first step in the translation phase is to parse the circuit description, and translate it into internal data structures. Once this has been done, the circuit is levelized, the gates of the circuit are sorted into levelized order, and each gate is simulated once to generate a set of consistent values. Next, a SIMULATION fan-out branch is added to each net visible to the user. Finally, a data structure known as a shadow is generated for each fan-out branch of each net in the circuit.

Fig. 9 illustrates the structure of a shadow. The **next** and **previous shadow** fields are used to link the shadow into the event list. The event list is doubly linked to facilitate dequeueing of events. The **subroutine** field points to the event processing routine for this fan-out branch. The **first** and **last fan-out branch** fields contain pointers to the first and last shadow that will be scheduled when an event propagates through the gate. All fan-out branches of a net are scheduled and descheduled simultaneously. To make this process more efficient, all shadows for a net are statically linked during the translation phase. This allows all shadows for a net to be inserted into the queue or deleted from the queue as a unit.

The **lock address** field contains the address of the dominant count for the gate associated with the fan-out branch. For NOT, BUFFER, XOR, and XNOR gates, this field is unused. Finally, the **queue address** identifies the queue into which the shadow is to be inserted.

Eight different event processors are used during the simulation phase of the inversion algorithm. These occur in pairs, and are called INCREMENT, INCREMENTX, DECREMENT, DECREMENTX, NOT, NOTX, XOR, and XORX. The second routine of each pair is used for shadows that are at the end of a subchain, while the second is used for the other shadows. The two subroutines of each pair are identical, except that the second routine removes the subchain from the queue. The routines were created in pairs to allow dequeueing to be performed without a conditional test. These routines are more detailed versions of the algorithms presented in Figs. 3 and 6.

We have created several different implementations of the inversion algorithm. Most of these use the zero-delay timing model, and are based on the LECSIM simulator developed by Wang [13]. (Unit-delay implementations of the inversion algorithm exist, but are beyond the scope of this paper.) LECSIM is a zero-delay event-driven levelized compiled code simulator. In LECSIM, gates are levelized and a queue is created for each level in the circuit, including the zero level. When a gate is queued for simulation, it is placed in the queue that corresponds to its level. Queues are processed in order by level. For asynchronous cyclic circuits, queues may be processed more than once.

Fig. 10. Structure of the simulation queue.

Like LECSIM, the zero-delay inversion algorithm levelizes the circuit and creates one event queue per level. Each queue consists of a doubly linked list of shadows terminated by a special shadow known as the queue trailer. The queue trailer is responsible for advancing the simulation from one queue to the next, and for terminating the simulation when appropriate. Fig. 10 illustrates the structure of the queue. As Fig. 10 illustrates, the queue headers are organized as an array of pointers, each of which points to a doubly linked list of shadows.

The simulation of an input vector begins with the primary input tests. The value of each primary input is compared with the value from the previous vector (or with zero for the first vector), and if there is a change, the fan-out branches of the primary input are inserted into queue zero. Once all primary input tests are complete, the simulator loads the address of the first shadow in queue zero into the **current-shadow** register and branches to the subroutine address contained in the shadow.

When an event is processed, additional shadows may be inserted into other queues. Once the last queue has been processed, simulation of the current vector terminates, and a new vector is read. Prior to reading the new vector, the value of each net visible to the user is printed. Fig. 11 gives the code for the INCREMENTX routine.

IV. Optimizations of the Inversion Algorithm

There are several simple optimizations that can significantly increase the performance of the inversion algorithm. The most important of these are the elimination of NOT and BUFFER gates, the elimination of XOR and XNOR gates, and the collapsing of homogeneous and heterogeneous connections.

A. Elimination of NOT and BUFFER Gates

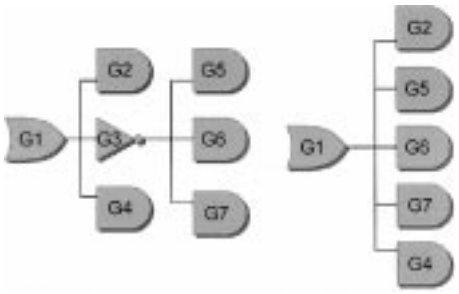
As the event-handler of Fig. 3 illustrates, the processing of NOT and BUFFER gates is a no-op operation in the inversion algorithm. When the input of a NOT or a BUFFER gate is processed, the only action that is taken is scheduling the fan-out branches of the output of the gate. The same simulation result can be achieved by eliminating the scheduling of NOT and BUFFER inputs and scheduling their fan-out branches instead. Fig. 12 illustrates this procedure. For the unit-delay and multidelay timing models, special procedures are required to preserve the delay of the gate.

B. Elimination of XOR and XNOR Gates

It is also possible to eliminate all XOR and XNOR gates. As with NOT's and BUFFER's, the only action taken when processing the input of an XOR or XNOR is scheduling or descheduling the fan-out branches of the gate. One can eliminate the processing of the input branch by scheduling or descheduling the output branches of the gate

```
INCREMENTX:
    Current_Shadow->subroutine = &DECREMENTX;
    (*Current_Shadow->Lock)++;
    if ((*Current_Shadow->Lock) == 1)
    {
        if (Current_Shadow->first_fanout->next == NULL)
        {
            Current_Shadow->last_fanout->next =
                Current_Shadow->Queue->next;
            Current_Shadow->Queue->next->previous =
                Current_Shadow->last_fanout;
            Current_Shadow->Queue->next =
                Current_Shadow->first_fanout;
            Current_Shadow->first_fanout->previous =
                Current_Shadow->Queue;
        }
        else
        {
            Current_Shadow->last_fanout->next->previous =
                Current_Shadow->first_fanout->previous;
            Current_Shadow->first_fanout->previous->next=
                Current_Shadow->last_fanout->next;
            Current_Shadow->last_fanout->next = NULL;
        }
    }
    Temp = Current_Shadow->next;
    Current_Shadow->next = NULL;
    Current_Shadow = Temp;
    Goto *Current_Shadow->subroutine;
```

Fig. 11. INCREMENTX event handler.



(a) (b)

Fig. 12. Elimination of a NOT gate. (a) Before elimination. (b) After elimination.

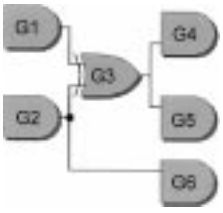


Fig. 13. Elimination of XOR gates.

instead. This can interfere with block scheduling of fan-out branches, as Fig. 13 illustrates.

Suppose that gate $G3$ of Fig. 13 has been eliminated, and that events propagate through both $G1$ and $G2$. Assume that the event for $G1$ is processed first. When the event for $G1$ is processed, it is necessary to schedule events for the inputs of $G4$ and $G5$. When the event for $G2$ is processed, it is necessary to *deschedule* the events for the inputs of $G4$ and $G5$, and schedule an event for the input of $G6$. If events are to be collapsed properly, the fan-out branches

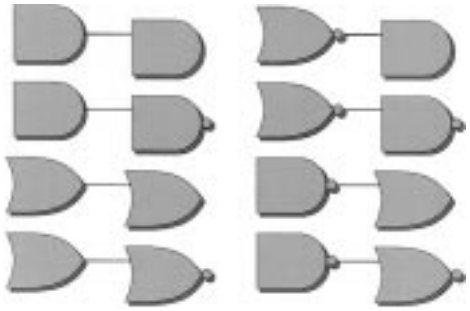


Fig. 14. Homogeneous connections.

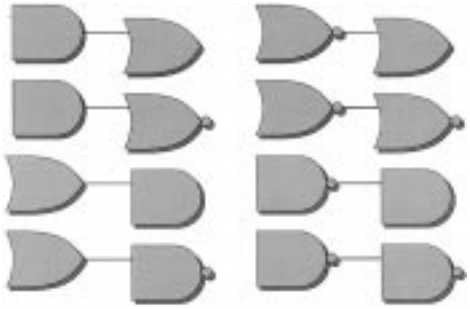


Fig. 15. Heterogeneous connections.

of $G3$ must maintain their identity, and cannot be grouped with the fan-out branches of $G1$ and $G2$. This problem arises only if one or more XOR input nets fan out to other gates. One could simply ignore event collapsing in these situations, but because of the relative rarity of XOR and XNOR gates, XOR/XNOR elimination has not been implemented in any current realization of the inversion algorithm.

C. Homogeneous and Heterogeneous Connections

Once all NOT, BUFFER, XOR, and XNOR gates have been eliminated from the circuit, all fan-out branches other than primary inputs and outputs must be connections among AND, OR, NAND, and NOR gates. These types of connections can be further categorized as *homogeneous* and *heterogeneous* connections. To distinguish between the two, suppose that net A is the output of $G1$ and the input of $G2$, and suppose that an event on the input of $G1$ propagates to $\boldsymbol{A}$. This will cause the dominant counts of both gates to change. If both counts are incremented or both are decremented, then the connection is homogeneous; otherwise, it is heterogeneous. Because a net may fan out to different types of gates, the heterogeneous and homogeneous properties apply to fan-out branches rather than to entire nets. It is possible to categorize connections at compile time using the tables illustrated in Figs. 14 and 15. When using these tables, it is necessary to do the categorization *before* NOT gates are eliminated. An intervening NOT gate changes a heterogeneous connection to a homogeneous connection, and vice versa. Two consecutive NOT gates cancel one another. A connection that passes through an XOR/XNOR gate cannot be categorized as either heterogeneous or homogeneous because the dominant counts of the two gates will sometimes move in the same direction and sometimes move in opposite directions.

D. Elimination of Homogeneous Connections

It is possible to eliminate all homogeneous connections from a circuit using the procedure illustrated in Fig. 16. To eliminate the

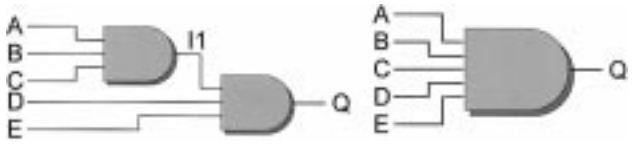


Fig. 16. Eliminating a homogeneous connection.

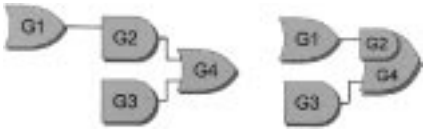


Fig. 17. Eliminating a heterogeneous connection.

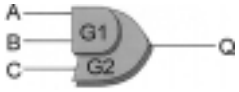


Fig. 18. Collapsed AND–OR connection.

connection $I1$, one simply removes gate $G1$, and treats the inputs A, B, and C as if they were inputs of $G2$. The initial value of the dominant count for $G2$ is recomputed by adding the initial dominant-counts of $G1$ and $G2$ and subtracting one.

The operation illustrated in Fig. 16 could be performed in a conventional simulator, but the benefits are not as clear. After the collapse, any event on inputs A, B, and C would result in the simulation of a five-input gate, regardless of whether the event would have propagated to $I1$ in the uncollapsed circuit. In the inversion algorithm, the processing of events on the inputs A, B, and C is identical in both circuits, and all processing for net $I1$ is eliminated in the collapsed circuit. Even if collapsing of heterogeneous connections proved to be beneficial in a conventional simulation, the ability to collapse connections is limited to those types illustrated in column 1 of Fig. 14. The connections in column 2 and connections with intervening NOT gates would pose a problem.

E. Eliminating Heterogeneous Connections

The procedures for eliminating heterogeneous connections are more complex than those for homogeneous connections, and do not always eliminate all operations for the connection. There are two procedures for eliminating heterogeneous connections: the *linear method,* and the *layered method.* The layered method allows more connections to be eliminated, but retains more operations for eliminated connections.

As Fig. 17 illustrates, the linear method can eliminate *only one* input from any gate. It is possible to collapse either $G2$ or $G3$ into $G4$. However, once $G2$ has been collapsed into $G4$, it is no longer possible to collapse $G3$ into $G4$. It is, however, possible to collapse gates in linear fashion by first collapsing $G1$ into $G2$, and then collapsing the $G1/G2$ combination into $G4$.

The linear method operates by changing the increment and decrement values used to update the dominant count of a gate. Instead of using a uniform value of 1, the linear method uses different values for different inputs. Consider the collapsed gate illustrated in Fig. 18.

In Fig. 18, a single dominant count will be maintained for the combined gate $G1/G2$. Suppose that all three inputs A, B, and C have the value 0. The dominant count corresponding to this input state is zero. For the output of $G1/G2$ to change, it is necessary for C to change or for *both* A and B to change. The total effect of both changes in A and B must equal the effect that $G1$ would have had in the original circuit, and must also equal the effect that C has on the collapsed gate. Neither A nor B by itself can have enough effect on the dominant-count to cause a change in the output of $G1/G2$. Because of the symmetry of the circuit, the effect of A and B must be the same.

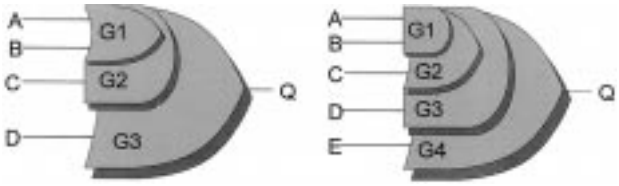


Fig. 19. Multilevel collapsed connections.

Fig. 20. Layered collapsing of connections.

There are many ways to assign increment/decrement values to the input nets A, B, and C that will achieve these requirements. One acceptable procedure is to assign the value 1 to the input C, and 0.5 to the inputs A and B. The output Q changes when the dominant count changes from a value less than 1 to a value greater than or equal to 1, or when it changes from a value greater than or equal to 1 to a value less than 1.

Collapsed connections with more than two levels follow the same principles as two-level collapsed connections. Values are assigned to inputs, depending on their relative power to change the output of the collapsed gate, and there are many different assignments that will achieve the desired results. Fig. 19 illustrates a three- and a four-level collapsed gate.

The values for the inputs of gate $G1/G2/3$ can be calculated in the following manner. Assume that A, B, C, and D have been initialized with the logic value of 0. The increment of D is set to 1. The total effect of the $G1/G2$ combination must be equal to the effect of D. Since A and B are symmetric inputs, the increments assigned to A and B should be equal. Since the output of $G2$ changes from 0 to 1 (thereby changing the output of $G1/G2/G3$) when C changes to 1 and either A or B changes to a 1, the sum of the increments assigned to A and C must equal 1. Since a change in both A and B, without an accompanying change in C, will not cause the output of $G1/G2/G3$ to change, it is necessary that the sum of the increments assigned to A and B be less than 1. This implies that the increment assigned to C must be greater than the increments assigned to A and B. To achieve these requirements, an increment of 0.25 is assigned to both A and B, while an increment of 0.75 is assigned to C. Using similar principles, the increments assigned to the inputs of $G1/G2/G3/G4$ are $A \rightarrow 0.125, B \rightarrow 0.125, C \rightarrow 0.25, D \rightarrow 0.75, E \rightarrow 1$. As with the two-level collapsed gate, the output changes value only when the gate count changes from a value less than 1 to a value greater than or equal to 1, or from a value greater than or equal to 1 to a value less than 1.

The layered method of collapsing heterogeneous connections allows arbitrary collapsing of connections, as illustrated in Fig. 20, but requires more computation in the simulation phase than the linear method.

```
INCREMENT_LAYERED_2:
Current_Shadow->subroutine = &DECREMENT_LAYERED_2;
(*Current_Shadow->Lock[1])++;
if ((*Current_Shadow->Lock[1]) == 1)
{
    (*Current_Shadow->Lock[0])--;
    if ((*Current_Shadow->Lock[0]) == 0)
    {
        if (fanouts not on queue)
        {
            insert fanouts into queue;
        }
        else
        {
            remove fanouts from queue;
        }
    }
}
Current_Shadow = Current_Shadow->next;
Goto *Current_Shadow->subroutine;
```

Fig. 21. INCREMENT routine for two-level connections.

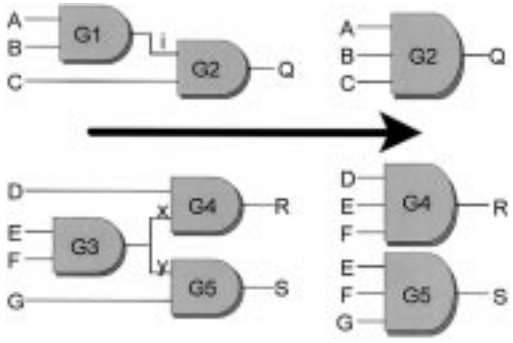


Fig. 22. Collapsing gates with fan-out.

Unlike the linear method, which uses a single dominant count for each gate, the layered method preserves the dominant counts of the original gates. Because the original dominant counts are preserved, the number of run-time increment/decrement operations is not reduced, nor is the number of tests reduced. However, the operations are performed hierarchically without any intermediate scheduling, which improves run-time performance. To illustrate, assume that in Fig. 20, input A of the collapsed gate changes from 0 to 1. The dominant count of $G1$ is decremented, and if the new value is not 0, no further processing is done. However, if the new value is 0, then the dominant count of $G3$ is incremented. If the new value is 1, then the fan-out branches of $G3$ are scheduled. No scheduling is done for the fan-out branches of $G1$ or $G2$.

The shadow of a layered connection differs from that illustrated in Fig. 9 in that the **Lock** component of the shadow is an array of pointers rather than a single pointer. Fig. 21 illustrates the code for a two-level layered connection. This code corresponds to the increment processor of a simple connection. As is the case for simple connections, both increment and decrement processors are used, the increment and decrement processors alternate with one another.

F. When to Collapse Connections

Although it is possible to collapse all homogeneous and heterogeneous connections in a circuit, it is not always advantageous to do so. Consider the two circuits illustrated in Fig. 22.

In the first circuit of Fig. 22, the net i does not fan out, so the connection can be collapsed by moving the inputs A and B of $G1$ to $G2$. In the second circuit, the output of $G3$ fans out into two branches x and y. To collapse this net, it is necessary to move the two inputs E and F to both $G4$ and $G5$. Because the inversion algorithm requires one event per fan-out branch, this doubles the number of events that must be processed when either E or F changes value. In the original circuit, a change in net E will cause one event to be processed, while in the collapsed circuit, two events will be processed. For the first circuit, a change in net A will cause one event to be processed in both the original and the collapsed circuit.

It is not immediately clear if the elimination of the events on x and y will offset the effect of doubling the inputs E and F. To further characterize the situation, the average number of events on the inputs and outputs of $G3$ were calculated, assuming all input vectors to be equally likely. For the uncollapsed connection, the average number of events on E, F, x, and y is 1.75 events per input vector. After collapsing the connection, the average number of events on E and F is 2.00 events per input vector. On average, the uncollapsed connection will be more efficient. Under the same assumptions, the average number of events on A, B, and i is 1.375 for the uncollapsed connection and 1.000 for the collapsed connection. In this case, the collapsed connection is more efficient.

Extending this analysis to gates with larger numbers of inputs and larger fan-outs, an uncollapsed two-input AND with a fan-out of 3 averages 2.125 events per vector, while the collapsed gate averages 3.00 events per vector. For a three-input AND with a fan-out of 2, the uncollapsed gate averages about 1.94 events per vector, while the collapsed gate averages 3.00 events per vector. A three-input AND with a fan-out of 3 gives even more striking results with an average of 2.16 events per vector for the uncollapsed gate and 4.50 events per vector for the collapsed connection. This analysis shows that it is advantageous to eliminate a connection only if it does not fan out to more than one gate.

V. Performance Evaluation

Four prototype simulators were constructed to test the performance of the inversion algorithm and the various optimizations discussed in the previous section. The first prototype is unoptimized; the second eliminates NOT and BUFFER gates; the third eliminates homogeneous connections, NOT gates, and BUFFER gates; and the fourth eliminates heterogeneous connections, homogeneous connections, NOT gates, and BUFFER gates. Heterogeneous connections were eliminated using the layered method. The linear method of eliminating heterogeneous connections was not tested. All prototypes are levelized event-driven zero-delay simulators based on the LECSIM model.

The ISCAS-85 benchmarks [15] were used to certify the correctness of the prototypes. Each circuit was simulated with 5000 randomly generated input vectors, and the outputs were compared to those of the FHDL simulator [20], a levelized compiled code simulator that has been in use for several years. These same circuits and input vectors were used to evaluate the relative performance of the four prototypes and the FHDL simulator. Each simulation was run on a SUN-4 IPC running SunOS with 12 Mbytes of memory and a dedicated disk drive. This system was isolated from outside influences as much as possible during the execution of the tests. To isolate the effects of each algorithm, each simulation was done three different ways. First, a complete simulation was done with full input and output. Next, the output functions of the simulators were disabled, leaving all input and simulation functions intact, and a second set of simulations was run. Finally, both the simulation functions and the output functions were disabled, leaving only the input functions intact, and a third set of simulations was run. Each of these simulations was performed five times, and the results were averaged to minimize errors in the UNIX /bin/time command, which

TABLE I
EXPERIMENTAL RESULTS

Circuit	Unopt.	NOT Elim.	Hom. Elim.	Hom/Het Elim.	LCC	Conv. Event-Drv	Activity Rate
c432	1.7	1.6	1.4	1.2	0.5	46.4	59.4
c499	2.0	1.9	1.9	1.9	0.6	51.1	63.2
c880	3.8	3.5	3.2	2.7	1.2	87.1	57.1
c1355	6.5	5.4	5.4	4.2	1.9	177.2	56.5
c1908	8.1	5.8	5.6	4.5	4.4	330.2	56.8
c2670	17.7	13.2	12.2	11.7	5.3	368.2	55.7
c3540	16.5	11.6	10.0	9.3	8.4	531.1	52.4
c5315	36.9	28.8	28.1	22.8	21.7	1024.0	63.8
c6288	40.4	40.0	39.7	33.8	30.1	9555.9	61.5
c7552	52.6	40.6	39.4	33.5	40.7	1483.2	60.7

was used to report the timings. The "user" field from the output of the /bin/time command was used to determine the execution time. The results of the read-only simulations were subtracted from the results of the no-print simulations to obtain the results reported in Table I. All simulations, except those for conventional event-driven simulations, were run as compiled code simulations. The C language was used as the target language for all of the simulators. No optimization flags were used when compiling the simulators.

As Table I indicates, the activity rates of the circuits tested ranged from just over 50% to over 60%. At this level of activity, levelized complied code simulation (the LCC column) typically outperforms event-driven simulation by a significant margin. However, for the inversion algorithm with deletion of homogeneous and heterogeneous connections, the timings are essentially the same for the circuits c1908, c3540, c5315, and c6288. For circuit c7552, the inversion algorithm actually outperforms levelized compiled code simulation. The performance of both the inversion algorithm and of conventional event-driven simulation are proportional to the activity rate. If the activity rate is reduced to a more reasonable level, the execution times of these algorithms will be reduced proportionally.

VI. CONCLUSION

As the results of the previous section indicate, the inversion algorithm is competitive with levelized compiled code simulation, even at very high activity rates. When the activity rate is reduced to a more reasonable level, the performance of the inversion algorithm will increase correspondingly, while the performance of LCC simulation will remain the same. Since the activity rates reported here are significantly higher than those that are likely to be encountered in practice, the inversion algorithm can be expected to outperform levelized compiled code simulation in most practical situations.

In addition to its performance, the inversion algorithm has several other desirable properties. The inversion algorithm can be run interpretively with only a minimal impact on performance. Therefore, it can be used for fast debugging of circuits during the initial phases of the design cycle, and for more massive testing later in the design cycle with only minor differences in performance.

Because of the small size of the run-time code, the inversion algorithm will be beneficial to tool developers who must support many different types of development platforms. The run-time code of the inversion algorithm can be written and debugged in a few hours, making it feasible to have several different assembly-language implementations of the same code.

The results shown here suggest that the inversion algorithm will be an effective tool for high-performance simulation of large circuits. In spite of this, there is a massive amount of research that must be done to realize the full potential of the techniques described in this paper. Work is currently underway to extend the inversion algorithm to more complex timing models and multiple-valued logic models. There is also work in progress to identify further optimizations, and to further characterize the underlying theoretical issues. The work described here should provide the foundation for much new research on the high-performance simulation of VLSI circuits.

REFERENCES

[1] R. E. Bryant, D. Beatty, K. Brace, K. Cho, and T. Sheffler, "COSMOS: A compiled simulator for MOS circuits," in *Proc. 24th Design Automation Conf.*, 1987, pp. 9–16.
[2] D. M. Lewis, "A hierarchical compiled code event-driven logic simulator," *IEEE Trans. Computer-Aided Design*, vol. 10, pp. 726–737, June 1991.
[3] M. Chiang and R. Palkovic, "LCC simulators speed development of synchronous hardware," *Comput. Design*, pp. 87–91, Mar. 1, 1986.
[4] W. Y. Au, D. Weise, and S. Seligman, "Automatic generation of compiled simulations through program specialization," in *Proc. 28th Design Automation Conf.*, 1991, pp. 205–210.
[5] A. W. Appel, "Simulating digital circuits with one bit per wire," *IEEE Trans. Computer-Aided Design*, vol. 7, pp. 987–993, Sept. 1988.
[6] C. Hansen, "Hardware logic simulation by compilation," in *Proc. 25th Design Automation Conf.*, 1988, pp. 712–715
[7] L. Wang, N. Hoover, E. Porter, and J. Zasio, "SSIM: A software levelized compiled-code simulator," in *Proc. 24th Design Automation Conf.*, 1984, pp. 473–478.
[8] Z. Barzilai, J. L. Carter, B. K. Rosen, and J. D. Rutledge, "HSS—A high speed simulator," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 601–617, July 1987.
[9] P. Maurer, "Two new techniques for unit-delay compiled simulation," *IEEE Trans. Computer-Aided Design*, vol. 11, pp. 1120–1130, Sept. 1992.
[10] Y. Lee and P. Maurer, "Two new techniques for compiled multi-delay simulation," in Proc. *SOUTHEASTCON'92*, Apr. 1992.
[11] ____, "Two new techniques for compiled multi-delay logic simulation," in *Proc. 29th Design Automation Conf.*, 1992, pp. 420–423.
[12] P. M. Maurer, "The shadow algorithm: A scheduling technique for both compiled and interpreted simulation," *IEEE Trans. Computer-Aided Design*, to be published.
[13] Z. Wang and P. M. Maurer, "LECSIM: A levelized event driven compiled logic simulator," in *Proc. 27th Design Automation Conf.*, 1990, pp. 491–496.
[14] S. P. Smith, M. R. Mercer, and B. Brock, "Demand driven simulation: BACKSIM," in *Proc. 24th Design Automation Conf.*, 1987, pp. 181–187.
[15] F. Brglez, P. Pownall, and R. Hum, "Accelerated ATPG and fault grading via testability analysis," in *Proc. Int. Conf. Circuits Syst.*, 1985, pp. 695–698.
[16] D. Schuler, "Simulation of NAND logic," in *Proc. COMPCON'72*, Sept. 1972, pp. 243–245.
[17] M. Breuer and A. Friedman, *Diagnosis and Reliable Design of Digital Systems.* Rockville, MD: Computer Science, 1976.
[18] M. Abramovici, M. Breuer, and A. Friedman, *Digital Systems Testing and Testable Design.* New York: Computer Science Press, 1990.
[19] P. Maurer and W. Schilp, "Three-valued simulation with the inversion algorithm," Dept. Comput. Sci. Eng., Univ. South Florida, Tampa, Tech. Rep. DA-27, 1995.
[20] P. Maurer, Z. Wang, C. Morency, A. Tokuta, and N. Bhate, "The Florida hardware design language," in *Proc. SOUTHEASTCON'90*, pp. 430–434.