Layout-driven Timing Optimization by Generalized De Morgan Transform

Supratik Chakraborty
Dept. of Computer Science and Engineering
Indian Institute of Technology, Bombay, India
supratik@cse.iitb.ac.in

Rajeev Murgai
Fujitsu Laboratories of America, Inc.
Sunnyvale, CA 94086, USA
murgai@fla.fujitsu.com

Abstract

We propose a timing-oriented logic optimization technique called Generalized De Morgan (GDM) transform, that integrates gate resizing, net buffering and De Morgan transformation. The contribution of our work lies in the integration of the three techniques, allowing them to interact at a much finer level of granularity than would be otherwise possible. This produces better results than those obtainable by individual techniques like net buffering or gate resizing applied to the circuit in various combinations. GDM transform is also layout-friendly since it does not alter the routing patterns and placement of cells, except possibly some buffer insertions/deletions. Hence it is useful for achieving timing closure in late stages of the design flow. We propose a comprehensive GDM algorithm that (a) determines the best replacement of a gate, possibly with inverted inputs and outputs, along with the best buffering configurations of nets incident on it, and (b) embeds this into a global scheme for optimizing large designs. We have implemented this algorithm in a layout-driven, industrial-strength logic optimization framework, and have successfully applied it to large industrial designs.

1. Introduction

Timing convergence is an important problem in deep submicron digital design. To address this problem, logic synthesis and technology mapping tools need access to layout information such as cell placements, wire lengths and wire delays, implying a tight coupling of physical and logical design flows. It is now standard practice to apply simple transformations such as gate resizing, net buffering, gate replication, etc. at design stages close to the layout process in order to achieve timing closure. These transformations, being simple, incremental and local in nature, do not change the layout significantly and are suitable for application immediately prior to detailed place-and-route. We have developed a tool for timing-driven logic optimization of digital circuits, suitable for application after global place-and-route but before detailed place-and-route. In this paper, we describe a technique, called *Generalized De Morgan* transform, implemented as part of this tool.

The input to our optimization tool is a circuit optimized by designers or logic synthesis tools and mapped to cells from a given library. It is assumed that a mincut based algorithm has been used to partition the chip area into blocks using horizontal and vertical cuts, and each cell is assigned to some block. A block may however contain multiple cells. Further, we assume that a global routing tool has been used to determine the

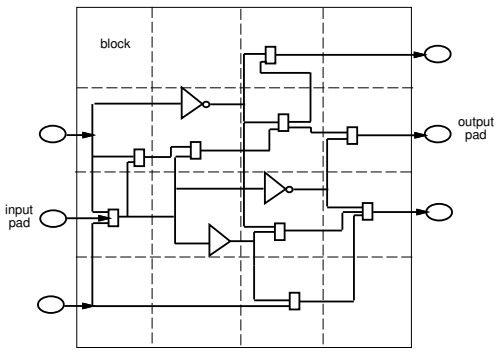

Figure 1. Block layout of a design

topology and global route of each net by generating net segments between blocks (see Figure 1). We argue that this is an appropriate stage in the design flow for applying optimization techniques such as net buffering and gate-resizing. Since cells have been roughly placed and nets have been globally routed, wire loads and delays are approximately known, and so is the delay at each pin. However, since the final layout has not been frozen, any optimization obtained through careful addition, deletion or replacement of cells can be incorporated into the final layout by refined placement and detailed routing. Fig. 2 shows our overall design flow and the position of the layout-driven optimization tools in the flow.

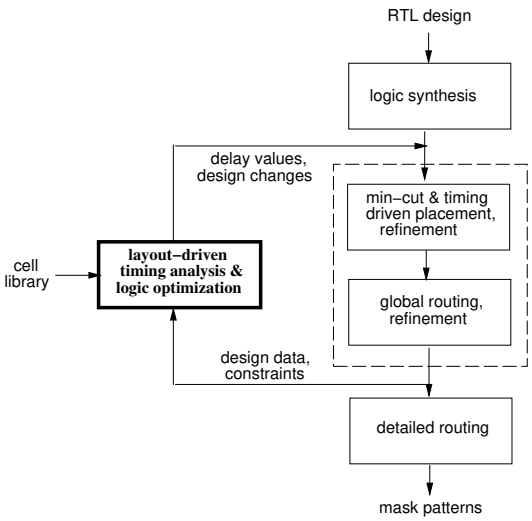

Figure 2. Design flow

Earlier, we had implemented well-known transformations such as gate resizing and net buffering in our tool. In this pa-

per, we propose a new technique called Generalized De Morgan (GDM) transform. This is more general than gate resizing and net buffering in that it changes the logic functionality of cells in the design. It is also more powerful than the simple De Morgan transform that replaces a gate with its dual (such as an AND with an OR) and inserts inverters at the inputs and outputs. As will be seen later, GDM transform integrates gate resizing, De Morgan's laws, and net buffering in one unified transform, allowing these three techniques to interact at a much finer level of granularity than would be otherwise possible. The transform is also layout-friendly: it replaces a gate with another one that has the same number of terminals, and preserves routing patterns and placement positions in the design (except for some buffer insertions/deletions). Thus, although GDM transform is less powerful than re-synthesis and re-mapping, it is more useful in a layout-driven scenario, especially in late stages of the design flow, where we wish to maximize circuit performance by minimal layout modifications.

In this paper, we propose a comprehensive GDM transformation algorithm that not only determines the best replacement of a cell along with the best buffering configurations of nets incident on it, but also embeds this into a global scheme for minimizing the delay of large designs. We have successfully applied this transform on large industrial designs built on state-of-the-art industrial libraries. Experiments indicate that one invocation of the GDM transform reduces the circuit delay more than can be obtained by net buffering alone, gate resizing alone, or by a combination of net buffering followed by gate resizing. The GDM transform applied in concert with net buffering and gate resizing yields the best possible delay results of all combinations of transformations in our repository.

2. Previous work

2.1. Simple transforms

Simple transforms are those that preserve the logic functionality of each cell in the design. Buffers and inverters may be added or deleted in these transforms. Gate resizing and buffer optimization constitute examples of simple transforms.

Gate Resizing: There are several approaches to gate resizing. These include transistor sizing approaches that model area and delay as posynomials [9], linear programming approaches using a piece-wise linear delay model [4], convex programming approaches [18], cutset based approaches [19] and global sizing [7], which is a greedy technique enhanced with a mechanism to come out of local minima.

Buffer Optimization: Existing work in this category falls in one of two domains – pre-routing and post-routing. In the pre-routing domain, the problem consists of generating a delay-optimum fanout tree and is known to be computationally hard. Several algorithms have been proposed targeting either a single multi-fanout gate and/or an entire design [10, 5, 22, 21, 13, 16]. In the post-global-routing realm, the topology of each net is fixed by the routing tool and the buffer optimization problem reduces to net buffering. Given a net with fixed topology and candidate buffer insertion points, van Ginneken proposed an optimum, polynomial-time, minimum-delay buffer insertion algorithm [15]. Extensions to handle an input-slew based delay model [14], optimal wire segmenting [2], noise [3] and resistive shielding effects [1] have also been proposed.

2.2. Complex transforms

More complex logic transformations may also be applied to a mapped circuit, resulting in the logic function of one or more cells to be changed. Examples of such transforms are De Morgan transform, and general re-synthesis and re-mapping.

In the LATTIS system [8], a number of optimizations such as gate resizing, buffer insertion, De Morgan, timing directed factorization and re-mapping are applied sequentially on a technology mapped netlist. The De Morgan transform used in LATTIS replaces a gate with its dual, and adds inverters at the inputs and outputs (if the complemented signal is not available). As will be shortly seen, the GDM transform proposed in this paper is more general than the De Morgan transform in LATTIS. Kannan *et al.* [12] and Ishioka [11] proposed post-placement, pre-routing optimizations, followed by incremental placement to incorporate the changes. Their transformations include fanout optimization and gate resizing applied sequentially [12], and re-synthesis and re-mapping [11]. In [12], the wire-routes are estimated by minimum spanning trees. Hojat and Villarubia [17] described the results of using simple synthesis operations such as gate sizing between partitioning steps in a mincut based placement algorithm. More recently, Shenoy *et al.* [20] proposed a design flow that incorporates an intermediate iterative step between the traditional synthesis and physical design steps. This step integrates a number of synthesis optimizations in the inner loop of an iterative method to perform global placement under given timing constraints.

A common drawback of all the above approaches is that the optimizations are performed before *true* routing information is available. Hence the assumptions regarding global net topology and wire loading and delays, which are crucial in deep submicron technologies, may be erroneous.

In [6], Carragher *et al.* proposed a paradigm in which logic optimization is interleaved with placement/partitioning refinement and hierarchical global routing. The optimizations incorporate a comprehensive set of layout-friendly, logic-level transformations for improving the delay and area of a mapped, block-placed, and globally routed circuit under design and technology constraints.

3. Preliminaries

Structurally, a *circuit* is an interconnection of cells through nets. An *extended net* is one that passes over buffers and inverters. The source of such a net is either a primary input or the output of a non-buffer and non-inverter cell. Sinks are either primary outputs or inputs of non-buffer and non-inverter cells. For example, let the boxes in Fig. 3a represent non-inverter and non-buffer cells. Then this subcircuit contains three extended nets: (i) with root G and sinks E and F, (ii) with root A and sinks C and I, and (iii) with root B and sinks J and D. Extended nets are the basic objects used in buffering algorithms. Since the input to our optimization tool is a fully-mapped, block-placed, and globally-routed design (as shown in Figure 1), the topology of extended nets are assumed to be known.

Timing analysis is done statically using cell delays, wire delays and wire loads. Pin-to-pin delays of cells can be computed either by the *load-dependent simple delay model* or *input-slew dependent linear delay model*. In the simple delay model, the delay from input pin i to output pin o of a cell is given by

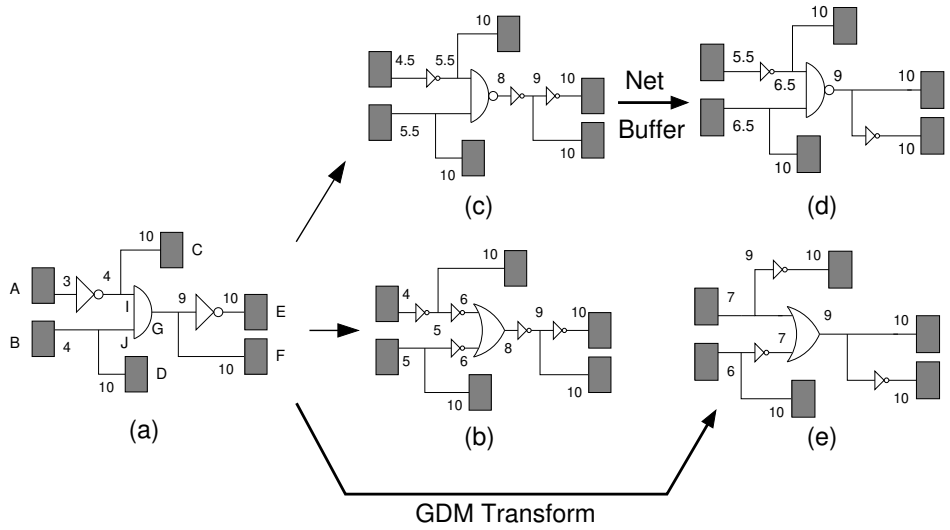

Figure 3. Comparison of GDM transformation with simple approach.

$d_{i,o} = \alpha_{i,o} + \beta_{i,o}.C_o$, where $\alpha_{i,o}$ is the intrinsic delay from i to o, $\beta_{i,o}$ is the load coefficient and C_o is the load capacitance visible at output o of the cell. The linear delay model has additional linear terms incorporating the slew at the input pin. Since we assume that global routing has been performed, approximate values of wire loads and delays can be derived. In this work, we use the Elmore delay model for computing wire delays.

The *arrival time* at a net is the time at which a signal propagating from the primary inputs reaches the net. The *required time* is the time by which the signal must reach the net for the circuit to meet timing requirements. The *slack* is the difference between the required time and the arrival time. A *critical* net is one that has the minimum slack in the circuit. A cell with a critical output net is a critical cell.

Given a globally routed extended net with required times and desired signal phases (even/odd inversions from source) at all sink nodes, an optimal net buffering algorithm chooses buffers and inverters from a library and inserts them at appropriate places on the extended net, such that all sink phase requirements are satisfied and the required time at the root is maximized. However, if the root is an output pin of a cell, the algorithm must not stop at maximizing the slack at the root; instead it must choose the buffering solution that maximizes the minimum slack at the inputs of the cell. In this paper, we assume that an algorithm similar to van Ginneken's buffering algorithm [15] (or its extension by Lillis *et al.* [14]) is available to us for this purpose.

4. GDM-equivalence of functions

Two boolean functions, f and g, are said to be *NN-equivalent* or *GDM-equivalent* if (a) both have the same support set, and (b) either f or f' can be obtained by complementing zero or more inputs of g. For example, let $f(x_1, x_2) = x_1 + x_2$ and $g(x_1, x_2) = x_1 x_2$. Since $f'(x_1, x_2) = g(x_1', x_2')$, functions f and g are GDM-equivalent.

More generally, let $\mathcal{F}(x_1, x_2 \ldots x_n)$ be a set of m functions $\{f_1, f_2 \ldots f_m\}$. The support set of each f_i is assumed to be a subset of $\{x_1, x_2 \ldots x_n\}$. Let I be an $(n+m)$-dimensional binary vector that encodes the complementation of inputs and outputs of $\mathcal{F}$ as follows. If the i^{th} bit of I (henceforth $I[i]$) is 1 and $1 \leq i \leq n$, then input x_i of $\mathcal{F}$ is complemented; otherwise, it is left uncomplemented. Similarly, if $I[n+i]$ is 1 and $1 \leq i \leq m$, the i^{th} output, f_i, of $\mathcal{F}$ is complemented; otherwise it is not complemented. Vector I is called the *inversion vector* of F, and the set of functions obtained after complementing inputs and outputs of F according to I is denoted $\mathcal{F}_I$. As an example, let $\mathcal{F}(x_1, x_2) = \{x_1 x_2, x_1\}$ and $I = [0, 1, 1, 0]$. Then $\mathcal{F}_I(x_1, x_2) = \{(x_1 x_2')', x_1\}$ or $\{x_1' + x_2, x_1\}$.

Using the above notation, two sets of functions $\mathcal{F}(x_1 \ldots x_n)$ and $\mathcal{G}(x_1 \ldots x_n)$ are said to be GDM-equivalent if both have the same number (m) of functions, and there exists an inversion vector I of $\mathcal{G}$ such that the i^{th} function of $\mathcal{G}_I(x_1 \ldots x_n)$ is equivalent to f_i for all i in 1 to m.

5. GDM transformation

GDM transformation of a cell F (multi-output in general) involves replacing F by a GDM-equivalent cell G, and inserting and/or deleting buffers and inverters on the extended nets at the inputs and outputs of G to preserve the functionality of F. In a simplistic approach, one would use the capacitance and delay information from the current buffering scheme (recall that we start with a mapped, globally placed and routed design) to determine the best GDM-equivalent cell $\mathcal{G}$ for replacing $\mathcal{F}$. Thereafter, inverters would be added at the appropriate inputs and outputs of $\mathcal{G}$ to make it functionally equivalent to $\mathcal{F}$. Finally, optimal net buffering would be applied to these nets to determine the best distribution of buffers and inverters, for the already determined choice of replacement cell $\mathcal{G}$. The drawback of this approach is that it fails to take into account the effect (capacitance and delays) of the final net buffering solutions when determining which replacement cell $\mathcal{G}$ to use in place of $\mathcal{F}$. Consequently, the optimal combination of replacement cell and net buffering solutions might be missed. The innovation in our approach lies in applying optimal net buffering to the input and output nets of the cell to implement inversions mandated by each choice of replacement cell $\mathcal{G}$, and then using this information to choose the best combination of $\mathcal{G}$ and input and output buffering. Thus, information from the final buffering scheme is used to guide the best choice of $\mathcal{G}$. While it may appear that computing buffering solutions for each input and output extended net for each choice of

$\mathcal{G}$ is prohibitively expensive, the computation of buffering solutions can be significantly optimized, as described in Section 5.3.

To illustrate the advantage of the GDM transformation vis-a-vis choosing replacement cell $\mathcal{G}$ first and then buffering input and output nets, consider the subcircuit shown in Fig 3a. Here, each shaded box represents a non-inverter/non-buffer cell sharing an extended net with the AND gate. The numbers represent required times at the corresponding locations in the circuit. We wish to replace the AND gate with a GDM-equivalent cell and buffer its inputs and outputs such that the minimum required time at A and B is maximized. For simplicity, let us assume that the cell library has only 4 cells: inverters with a fixed delay of 1, two-input AND gates with a fixed delay of 5 from each input to output, two-input NAND gates with a fixed delay of 2.5 from each input to output, and two-input OR gates with a fixed delay of 2 from each input to output. The variation in the delays of AND, NAND and OR gates could be attributed to different gate sizes, for example.

If we choose to replace the AND gate first, and then insert inverters at its inputs and/or outputs, we have two choices as shown in Figs. 3b and c. Since the minimum required time at A and B in Fig. 3c is higher than that in Fig. 3b, the NAND gate is chosen as the replacement for the AND gate. Subsequently, on applying net buffering techniques to the input and output nets of the NAND gate, the subcircuit in Fig. 3d results. Here, the minimum required time at A and B is 5.5. On the other hand, if we apply GDM transformation, the subcircuit shown in Fig. 3e is obtained. The minimum required time at A and B in this subcircuit is 6; hence this is a better transformation than Fig. 3d. Since GDM transformation considers the effect of buffering input and output nets when determining the best choice for the replacement cell, it is able to determine that the subcircuit in Fig. 3b eventually leads to a better subcircuit (Fig. 3e) than the subcircuit eventually obtainable from Fig. 3c.

It is important to observe that gate sizing, net buffering, and simple De Morgan transform (i.e., replacing a gate with its dual, such as an AND with an OR) can be viewed as special cases of GDM transform. For instance, gate sizing is subsumed in the GDM transform, since a function f is GDM-equivalent to itself (by definition) and hence GDM transformation automatically considers replacing a cell $\mathcal{F}$ by another cell $\mathcal{G}$ that has the same functionality as $\mathcal{F}$, but has a different size. However, our current GDM implementation also imposes some restrictions, as discussed in the next section.

5.1. Basic algorithm

The basic algorithm for choosing the best GDM transform of a cell F for timing optimization is shown in Fig 4. Inverters and buffers are not referred to as cells in this discussion. Function GDM_cell does not really implement any transformation; it simply determines the transformation that maximizes the minimum slack at the inputs of cells driving F (henceforth called local minimum slack). The procedure for applying GDM transform globally to the entire circuit is divided into 3 phases, as described below.

***(a) Evaluation*:** Let C be the set of all critical cells in the circuit. For each cell F in C, the best GDM transformation is determined by invoking function GDM_cell. Each transformation (G^*, N^*) thus obtained is assigned a cost, based on (a) local

GDM_cell (circuit K, cell F, library L)
In = set of cells/pr. inps. driving extended input nets of F.
Out = set of cells/pr. outs driven by extended output nets of F.
s_F = minimum slack at input pins of cells in In.
E = set of cells in L that are GDM-equivalent to F.
1. Save current buffering of input & output extended nets of F.
2. For each cell G in E
   (a) Replace F by G
   (b) I = inversion vector to make G equivalent to F.
   (c) Apply optimal net buffering routine to input & output extended nets of G to implement inversions in I.
   (d) Propagate required times from cells in Out to those in In using parameters of G, and buffering of step 2(c)
   (e) $s_G \leftarrow$ new min slack at the inputs of cells in In
3. Restore F and original buffering of extended nets.
4. Let G^* be the replacement cell that maximizes s_G, and N^* be the corresponding buffering solutions.
5. If $s_{G^*} > s_F$, return (G^*, N^*)
   else return (F, original net buffering).

Figure 4. Finding best GDM transform of a cell.

min slack improvement $(s_{G^*} - s_F)$ in function GDM_cell, and (b) area penalty of replacing the original cell F by G^* and due to implementation of buffering solution N^*. Let T be the set of best GDM transformations for all cells in C. The set T is kept sorted in increasing order of the cost of transformations.

***(b) Selective implementation*:** Given the sorted set T, the transformation with the least cost is first implemented. Let this be the transformation (G^*, N^*) for critical cell F in the original circuit. Now, implementing N^* potentially modifies the buffering of some input and output extended nets of F. Therefore, if there exists another critical cell F' that shares an input or output extended net with F, then there can be a potential conflict between the buffering solutions for the shared net in the GDM transformations for F and F'. Thus, once the GDM transformation for F is implemented, the already-computed GDM transformation for F' may be rendered sub-optimal or even invalid (e.g., if they differ in the phases of signals at input or output pins of G^*). Therefore, the GDM transformations of all critical cells that share an extended net with F are removed from the set T after (G^*, N^*) is implemented. Of the remaining transformations in T, the one with the lowest cost is then chosen for implementation and the process is repeated until the set T is empty.

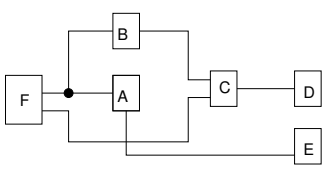

Figure 5. Illustrating sharing of extended nets.

As an example, suppose cells A, B, C, D, E and F in Fig 5 are critical. Initially, the best GDM transformation for all cells are determined. Suppose the cost of transformation increases from A to F. Using the above strategy, once the transformation for A is implemented, those for B, E and F are invalidated because they share an extended net with A. The transformation for C is thus implemented next. This, in turn, invalidates the transformation for D. Therefore, only A and C are trans-

formed. A consequence of this invalidation policy is that the current GDM implementation does not strictly subsume our gate resizing or net buffering implementation. Our gate resizing algorithm can potentially resize any gate, whereas a given iteration of GDM will not simultaneously resize two gates that share extended nets. Similarly, GDM transformation may not buffer two extended nets N_1 and N_2 simultaneously if the root cells C_1 and C_2 of N_1 and N_2 share a net between them, and C_1 is GDM transformed.

***(c) Partial undoing*:** Once the set T becomes empty, a static timing analyzer is run on the modified circuit to determine the new minimum slack. If this is larger than the original minimum slack s_0 (before the transformations were applied), the new set C of critical cells is determined and phases (a), (b) and (c) are repeated. However, if the minimum slack worsens as a result of the transformations[1], we undo the transformations in small groups in the following manner. We first undo a fixed small number (10 in our experiments) of the costliest transformations and re-evaluate the minimum slack in the circuit. If the minimum slack increases beyond the original slack s_0, we accept the remaining transformations (not yet undone), and repeat phases (a), (b) and (c). Otherwise, we undo a small number of the remaining costliest transformations, and continue the process until either the minimum slack increases beyond s_0 or all transformations are undone. If all transformations are undone, we terminate the iterative process.

Observe that in phase (b), one can choose to re-compute the best GDM transformations of all cells on a shared extended net whenever one of the gates is transformed. However, this entails significant computation every time a transformation is implemented, and most of the results re-computed are soon invalidated again. Moreover, given the inevitable inaccuracies in estimating wire delays and capacitances at this stage of the design flow, we believe that this is not worth the extra effort. Our experiments also suggest that even if the GDM transformation for a cell G is invalidated in an iteration, there is a good chance of G being reconsidered (after re-evaluation of its best GDM transformation) in a later iteration if G continues to remain critical.

While the basic algorithm for timing optimization of a circuit by GDM transformation has been given above, there are several important issues related to implementation. We discuss them briefly in the following subsections.

5.2. Equivalence classes of library cells

GDM-equivalence is easily seen to be reflexive, symmetric and transitive. Therefore, GDM-equivalence is an equivalence relation and can be used to partition the cells in a library into disjoint equivalence classes. The motivation for creating equivalence classes comes from the observation that once these classes are created, determining the set E in function GDM_cell (see Fig. 4) reduces to accessing the equivalence class of F. Since the number of critical cells in real designs often runs into thousands, it is beneficial to compute the equivalence classes once and reuse the results.

In order to determine all cells that are GDM-equivalent to $\mathcal{F}$, we proceed by filtering out as many cells as possible by a series of simple tests. These tests either filter out a cell $\mathcal{G}$, or yield information about values of bit positions in the inversion vector I of $\mathcal{G}$ for it to be GDM equivalent to $\mathcal{F}$. For the remaining bit positions, we need to assign all possible combinations of bit values to obtain the possible inversion vectors. For each vector I thus generated, we then use reduced ordered binary decision diagrams (ROBDDs) to check the equivalence of each function f_i in $\mathcal{F}$ with the corresponding function of $\mathcal{G}_I$. If an input vector I is obtained such that the BDD equivalence checks succeed for all functions f_i of $\mathcal{F}$, we put $\mathcal{G}$ in the same equivalence class as $\mathcal{F}$.

Below, we list some simple tests to eliminate a candidate cell G from the equivalence class of $\mathcal{F}$.

1. If the number of inputs and outputs of $\mathcal{G}$ and $\mathcal{F}$ don't match, $\mathcal{G}$ is not GDM-equivalent to $\mathcal{F}$.
2. Let there be n inputs and m outputs of each of $\mathcal{G}$ and $\mathcal{F}$. Let $N(g_i)$ be the onset count (number of minterms) of function g_i of $\mathcal{G}$ and let $N(f_i)$ be the onset count of the corresponding function f_i of $\mathcal{F}$ (correspondence is by index number). The onset counts can be easily determined once BDDs of the functions in $\mathcal{G}$ and $\mathcal{F}$ are constructed. If $N(g_i) \neq N(f_i)$, then g_i is not equivalent to f_i. In addition, if $N(g_i) \neq 2^n - N(f_i)$, then g_i is not equivalent to f_i' either. Therefore, if both tests are satisfied, $\mathcal{G}$ is not GDM-equivalent to $\mathcal{F}$.
3. Let $g_i|_a$ be the cofactor of function g_i of $\mathcal{G}$ with respect to input a. Using the notation for onset counts, If $N(g_i|_a) \neq N(f_i|_a)$ or $N(g_i|_{a'}) \neq N(f_i|_{a'})$, then g_i and f_i cannot be equivalent with the same phase of input a applied to both. Similarly, if $N(g_i|_a) \neq N(f_i|_{a'})$ or $N(g_i|_a) \neq N(f_i|_{a'})$, then g_i and f_i cannot be equivalent with opposite phases of input a applied to f_i and g_i. In a similar manner, we can check from the onset counts if it is possible for g_i and f_i' to be equivalent with same (opposite) phase of input a applied to both. If all the above tests fail (i.e., neither f_i nor f_i' can be equivalent to g_i), we infer that f_i is not GDM-equivalent to g_i; therefore $\mathcal{F}$ and $\mathcal{G}$ are not GDM-equivalent.

When the tests involving onset counts fail to eliminate $\mathcal{G}$ from the equivalence class of $\mathcal{F}$, they usually provide useful information about the inversion vector I of $\mathcal{G}$. For example, suppose $N(g_i) = N(f_i)$ and $N(g_i) \neq 2^n - N(f_i)$. It is easy to infer from this that if $\mathcal{G}$ and $\mathcal{F}$ are GDM-equivalent, the i^{th} output of each must have the same phase, i.e., the $n + i^{th}$ bit in the inversion vector must be 0. Now, suppose another test (using onset counts of cofactors), indicates that g_i and f_i must be in opposite phase if $\mathcal{F}$ and $\mathcal{G}$ are GDM-equivalent. Since we have a contradiction, we can then conclude that $\mathcal{G}$ is not GDM-equivalent to $\mathcal{F}$. In other cases, when no such contradictions arise, the information about bit values in the inversion vector helps reduce the number of BDD equivalence checks that need to be finally performed. For example, suppose in an 8-bit inversion vector I, we have already inferred the values of 4 bits from tests involving onset counts. In order to complete our test of whether $\mathcal{G}$ is GDM-equivalent to $\mathcal{F}$, we need to check if the BDD for each f_i in $\mathcal{F}$ is equivalent to the BDD for the corresponding function in $\mathcal{G}_I$ for one out of 2^4 combinations of unknown bit values in I. In contrast, if we did not have any knowledge of bit values in I, we would have needed to consider 2^8 combinations.

[1] Although each GDM transformation doesn't worsen the slack when considered in isolation, interactions between multiple transformations can cause a degradation.

Our method for identifying GDM-equivalence classes is able to identify interesting cases where a gate is GDM-equivalent to another gate of the same functionality, but with some inputs and/or outputs inverted. For example a two-input XOR is equivalent to another two-input XOR with either (i) no inputs or outputs inverted, or (ii) one input and output inverted, or (iii) both inputs inverted. Such equivalences cannot be identified by gate sizing or net-buffering alone, but can prove useful when trying to optimize a design with minimal layout changes.

5.3. Buffering extended nets at inputs and outputs

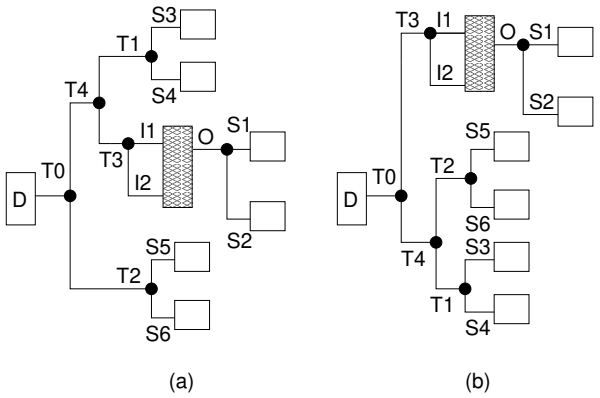

Figure 6. Buffering extended nets.

We saw in Section 5.1 that for each candidate cell $\mathcal{G}$ that is GDM-equivalent to $\mathcal{F}$, we must compute optimal net buffering solutions for extended nets at the inputs and outputs of $\mathcal{G}$. However, re-computing the solutions each time is very time consuming and wasteful. By carefully analyzing the problem, it is possible to store partial buffering solutions at suitable nodes on the extended nets and reuse them for all candidate cells $\mathcal{G}$. This leads to significant run-time improvements of the GDM transformation.

Let us first consider extended nets at cell outputs. As an example, suppose the shaded cell in Fig. 6 is being considered for GDM transformation. The extended net from output O goes to sink pins S_1 and S_2 of other cells. Clearly, GDM transformation of the shaded cell does not affect the required times or phases of signals at S_1 and S_2 under the load-dependent delay model. To see why the phases are not changed, recall that GDM transformation of a cell preserves the functionality of the cell, as viewed from other cells. The optimal buffering solution for this extended net, as computed by van Ginneken's [15] algorithm (with Lillis' [14] extensions) depends only on the required times and signal phases at the sink nodes, and hence is independent of the replacement cell used for the shaded cell. The buffering algorithm essentially generates two solutions: one which expects an inverted signal from the source of the net, and the other which expects the driving signal to be non-inverted. The best solution is then obtained by choosing the one that maximizes the required time at the source. This strategy for solving the buffering problem perfectly suits our requirements, since some GDM-equivalent cells require inversion of the output before driving the extended net, while others do not. Thus, instead of rejecting one solution, we can store the buffering solutions for both phases of the output at the root node O of the extended net. This essentially amounts to stopping short of the final step in van Ginneken's algorithm. Depending on whether an inversion is mandated on the output by the GDM-equivalent cell or not, we can then choose the appropriate solution from the stored pair of solutions. This effectively reduces the number of computations of buffering solutions for each extended net at the output to only one computation.

Let us now consider extended nets at the inputs of the cell. In Fig. 6a, the net from D to sink nodes I_1, I_2, S_3, S_4, S_5 and S_6 represents one such net. In this case, the required times and phases at sink nodes I_1 and I_2 can change as a result of GDM transformation of the shaded cell. However, the required times and phases of S_3, S_4, S_5 and S_6 do not depend on the transformation of the shaded cell. To compute buffering solutions for the input extended net efficiently, we need to examine van Ginneken's optimal net buffering algorithm [15] more closely. In his algorithm, solutions are first computed at each sink node of a net, and then they are propagated towards the root node by merging solutions from the child nodes at the parent Steiner node. So, in our example, solutions from S_3 and S_4 can be merged at Steiner node T_1, and similarly solutions from S_5 and S_6 can be merged at T_2, regardless of the cell used to replace the shaded cell. However, the buffering solutions from I_1 and I_2 depend on the required time and signal phases at I_1 and I_2, which in turn, depend on the replacement cell and the corresponding output buffering solutions. Therefore, we cannot propagate solutions from I_1 and I_2 to T_3 before choosing the replacement cell for the shaded cell. This, in turn, prevents solutions at T_1 from being propagated to T_4 and solutions at T_2 from being propagated to T_0. Thus, we cannot compute the buffering solution for the complete extended net independent of the choice of the replacement cell. Nevertheless, we can save some re-computation by storing partial results that do not depend on the choice of the replacement cell at intermediate nodes such as T_1 and T_2. These partial results can then be used to compute complete buffering solutions for the input extended nets for all candidate replacement cells. In our example, once a choice of the replacement cell and the corresponding output buffering solution is known, the required times and capacitances at nodes I_1 and I_2 are computed and the solutions propagated to T_3 and then to T_4 and finally to T_0. Note that the amount of computational savings in this case depends on the net topology. For example, if the extended net at the input had the topology shown in Fig. 6b, we could have propagated the solutions from S_3, S_4, S_5 and S_6 to T_4. Thus, we would have needed to perform 4 computation/merging of input buffering solutions (at I_1, I_2, T_3 and T_0) for every choice of the replacement cell. In contrast, the number of computation/merging for the corresponding net in Fig. 6a is 5 per choice of a replacement cell.

6. Experimental results

To test the effectiveness of the GDM transform, we used six optimized, mapped, block-placed, and globally routed industrial designs. Table 1 shows the relevant statistics of these designs such as the number of gates (a gate could be as simple as an inverter or as complex as an 8-bit adder), number of simple nets, the total gate area of the design (in terms of the smallest inverter in the library), and the original delay (i.e., delay before optimization by our tool) of the mapped, block-placed, and globally-routed circuit. This delay takes into consideration pin-to-pin delays through the cells using the input slew-based linear delay

Example	NB			GS			GS after NB			GDM			
	τ_{new} (ns)	ΔA (BC)	CPU (sec)	τ_{new} (ns)	ΔA (BC)	CPU (sec)	τ_{new} (ns)	ΔA (BC)	CPU (sec)	τ_{new} (ns)	ΔA (BC)	#invs	CPU (sec)
ex1	4.84	96	4	5.16	187	6	4.84	265	5	4.49	201	6	18
ex2	7.32	70	58	7.32	112	80	7.32	70	67	7.32	78	1	67
ex3	10.50	436	199	12.43	97	142	10.42	775	260	10.61	417	3	174
ex4	9.11	1075	214	9.03	676	202	9.11	1644	258	8.77	4653	61	392
ex5	10.50	622	485	12.13	1157	371	10.24	3552	673	9.70	3001	104	1133
ex6	43.00	2058	1444	48.62	583	1479	40.15	2604	1679	40.98	2508	48	2133

Example	GDM after NB after GS		
	τ_{new} (ns)	ΔA (BC)	CPU (sec)
ex1	4.49	352	16
ex2	7.32	70	73
ex3	10.42	775	288
ex4	8.75	1757	285
ex5	9.53	4302	934
ex6	39.79	3004	2222

NB: Net Buffer; GS: Gate Size; GDM = Generalized De Morgan transform.
τ_{new} = Delay after optimization; ΔA = Area Penalty.
1 BC = area of smallest inverter in library.
#invs = # cell replacements involving ≥ 1 inversion at input/output nets.
CPU time includes time to read circuit.

Table 2. Experimental Results

Example	#Gates	#Nets	Gate Area (BC)	Original Delay (ns)
ex1	356	409	1567	7.84
ex2	17.1K	26.0K	122.6K	7.44
ex3	32.0K	37.7K	343.6K	14.49
ex4	40.0K	48.1K	200.2K	11.38
ex5	86.7K	108.1K	381.6K	18.41
ex6	172.2K	210.9K	718.6K	56.36

1K = 1000, 1 BC = area of smallest inverter in library.
All benchmarks are in 0.35-μ technology.

Table 1. Benchmark Statistics

model, and the wire loads and delays using the Elmore delay model. The two largest designs, ex5 and ex6, are *hi-vision TV encoder/decoder* designs. ex6 has 172K gates and 211K simple nets. All experiments were performed on an UltraSparc 60 with 768MB RAM.

To evaluate the power of the GDM transform, we compared it with net buffering (NB), with gate resizing (GS), and with a composed transform consisting of net buffering followed by gate resizing (NB + GS). The results are shown in Table 2. It can be seen that GDM is much more effective in reducing the design delay than NB or GS on all designs, except ex2 on which it yields the same delay, and ex3 on which it is marginally worse than NB, although much better than GS. The worse performance of GDM as compared to NB can be attributed to the following fact. As mentioned in Section 5.1, our GDM implementation does not strictly subsume gate resizing or net buffering, because it prohibits simultaneous resizing of any gates that share nets or simultaneous buffering of two nets N_1 and N_2 whose root cells share either N_1 or N_2. The effectiveness of GDM vis-a-vis NB or GS can be correlated with the number of cell replacements which require inversions at the input or output nets of the cell (column #invs in Table 2). In the cases of ex2 and ex3, the number of such cells is very small: 1 and 3 respectively, and GDM is not very effective. However, on ex4, ex5, and ex6, this number is much higher, resulting in greater delay improvement by GDM vis-a-vis NB and GS.

We also find that GDM is more powerful than a combination of NB and GS in three out of six designs, equally powerful in one, and less powerful in the remaining two. Finally, to see the combined effect of all transforms, we applied NB followed by GS followed by GDM (we found empirically that this yields better results for our benchmarks than those obtained by applying GDM after NB after GS). It can be seen from Table 2 that this achieves better delay results than NB + GS or GDM alone. The average delay improvement over NB + GS is 3.2%, with the best improvement being 7% for ex1 and ex5. Although this may seem a small improvement, it is significant for the following reasons:

1. Net buffering and gate resizing are the most popular in-place optimization techniques used during layout in the industry today. However, if a circuit fails to meet timing requirements even after an application of these techniques, the designer is left with no choice but to painstakingly re-synthesize, re-map, re-place, and re-route parts of the design. As can be seen from Table 2, the GDM transform provides an additional in-place optimization technique that can come to the rescue of the designer. The extra 4-5% delay improvement obtained by GDM transform might just be enough to meet the timing requirements.

2. Since GDM transformation is applied to a design that is close to the final layout, even small timing improvements can translate to improvements in the final circuit delay.

Of course, there is a run-time penalty associated with GDM. As compared to NB + GS, GDM takes 66% more CPU time on average. Also, NB + GS + GDM takes 53% more time than NB + GS. We discuss one possible way to reduce the run-time in the next section.

7. Excessive buffering penalty

In some circuits, such as ex4 in Table 2, our current implementation of GDM can cause excessive area penalty. Consider

a cell $\mathcal{F}$ that is being GDM-transformed. Suppose the replacement cell for $\mathcal{F}$ is $\mathcal{G}$. The current implementation of the algorithm attempts to optimally insert buffers and/or inverters for reducing the delay on every input and output extended net of $\mathcal{G}$, regardless of whether the net is critical or not. Since delay optimization is usually area-hungry, this leads to a large number of buffers being inserted on several non-critical input and output nets and thus incurs a significant area penalty. Indeed, the buffering algorithm invoked on a net determines the minimum delay buffering strategy for the net regardless of whether it is critical or not. Therefore, it should be the onus of the GDM algorithm to present only critical nets to the delay-reducing buffering algorithm; the remaining nets should be buffered, if needed (for ensuring correct phases of signals at sink nodes), using an area-aware delay-constrained algorithm, such as in [14]. This optimization is currently being investigated in our tool. We expect the area penalty due to buffering of input and output nets to reduce considerably as a result of this optimization. We also expect a signficant speed-up in the run-time of the transformation, since we can skip computing buffering solutions of non-critical nets for which the source and sink phases remain unchanged.

8. Conclusions

In this paper, we have proposed a timing-oriented logic optimization technique, called Generalized De Morgan (GDM) transform, that allows gate resizing, net buffering and De Morgan transformation to interact at a fine level of granularity. It is a powerful transform for optimizing circuits in the post-global route-and-place, pre-detailed route-and-place stage of the design flow. At this stage, maximal improvements in performance with minimal changes to the circuit configuration and layout are desirable. The GDM transform replaces one cell by a GDM-equivalent cell with the same number of inputs and outputs, and simply inserts and/or deletes buffers and inverters on the extended input and output nets. Consequently, the placement and routing is minimally disturbed. It deserves mention here that there exist other techniques like transistor reordering that also attempt to minimize delay without significantly disturbing the layout. However, transistor reordering has not been considered in the current work.

Our experimental results indicate that one pass of GDM transformation applied to circuits in our benchmark suite produces delays that are better than those obtained by gate sizing and net buffering applied in concert. Although GDM is less powerful than arbitrary re-synthesis and re-mapping, it is more layout-friendly and is therefore of greater practical utility in a layout-driven framework. We believe that powerful yet layout-friendly transforms like the GDM transform are important steps in the search for a predictable and acceptable solution of the timing convergence problem.

References

[1] C. J. Alpert, A. Devagan, and S. T. Quay. Buffer insertion with accurate gate and interconnect delay computation. In *Proceedings of the Design Automation Conference*, pages 479–484, June 1999.

[2] C. J. Alpert and A. Devgan. Wire Segmenting For Improved Buffer Insertion. In *DAC*, pages 588–593, 1997.

[3] C. J. Alpert, A. Devgan, and S. T. Quay. Buffer Insertion for Noise and Delay Optimization. In *Proceedings of the Design Automation Conference*, pages 362–367, 1998.

[4] M. Berkelaar and J. Jess. Gate Sizing in MOS Digital Circuits with Linear Programming. In *The Proceedings of the European Conference on Design Automation*, 1990.

[5] C. L. Berman, J. L. Carter, and K. F. Day. The Fanout Problem: From Theory to Practice. In C. L. Seitz, editor, *Advanced Research in VLSI: Proceedings of the 1989 Decennial Caltech Conference*, pages 69–99. MIT Press, Mar. 1989.

[6] R. Carragher, R. Murgai, S. Chakraborty, M. Prasad, A. Srivastava, and N. Vemuri. Layout-driven logic optimization. In *Proceedings of the International Workshop on Logic Synthesis*, pages 270–276, June 2000.

[7] O. Coudert, R. Haddad, and S. Manne. New Algorithms for Gate Sizing: A Comparative Study. In *Proceedings of the Design Automation Conference*, pages 734–739, 1996.

[8] J. P. Fishburn. LATTIS: An Iterative Speedup Heuristic for Mapped Logic. In *29 ACM/IEEE Design Automation Conference*, pages 488–491, 1992.

[9] J. P. Fishburn and A. E. Dunlop. TILOS: A Posynomial Programming Approach to Transistor Sizing. In *Proceedings of the International Conference on Computer-Aided Design*, pages 326–328. IEEE, 1985.

[10] H. J. Hoover, M. M. Klawe, and N. J. Pippenger. Bounding Fanout in Logical Networks. *Journal of the Association for Computing Machinery*, 31(1):13–18, Jan. 1984.

[11] T. Ishioka, M. Murofushi, and M. Murakata. Layout Driven Delay Optimization With Logic Re-synthesis. In *Workshop Notes of the International Workshop on Logic Synthesis*, 1997.

[12] L. Kannan, P. Suaris, and H. G. Fang. A Methodology and Algorithms for Post-Placement Delay Optimization. In *Proceedings of the Design Automation Conference*, pages 327–332, 1994.

[13] D. Kung. A Fast Fanout Optimization Algorithm for Near-Continuous Buffer Libraries. In *Proceedings of the Design Automation Conference*, pages 352–355, 1998.

[14] J. Lillis, C. K. Cheng, and T. T. Y. Lin. Optimal Wire Sizing and Buffer Insertion for Low Power and a Generalized Delay Model. In *ICCAD*, pages 138–143, 1995.

[15] L. P. P. P. van Ginneken. Buffer Placement in Distributed RC-tree Networks for Minimum Elmore Delay. In *Proceedings of the International Symposium on Circuits and Systems*, pages 865–868, 1990.

[16] P. Rezvani, A. Ajami, M. Pedram, and H. Savoj. LEOPARD: A Logical Effort based Fanout OPtimizer for ARea and Delay. In *Proceedings of the International Conference on Computer-Aided Design*, pages 516–519, November 1999.

[17] S. Hojat and P. Villarubia. An Integrated Placement and Synthesis Approach for Timing Closure of PowerPC Microprocessor. In *International Conference on Computer Design*, pages 206–210, 1997.

[18] S. S. Sapatnekar, V. Rao, P. Vaidya, and S. Kang. An exact solution to the transistor sizing problem for cmos circuits using convex optimization. *IEEE Transactions on Computer-Aided Design*, CAD-6(6):1621–1634, Nov. 1993.

[19] H. Savoj, K. Xiang, K. Pan, and A. Domic. Technology dependent timing optimization. In *Proceedings of the International Workshop on Logic Synthesis*, June 1997.

[20] N. Shenoy, M. Iyer, R. Damiano, K. Harer, H.-K. Ma, and P. Thilking. A Robust Solution to the Timing Convergence Problem in High Performance Designs. In *International Conference on Computer Design*, October 1999.

[21] K. J. Singh. *Performance Optimization of Digital Circuits*. PhD thesis, UC Berkeley, Dec. 1992.

[22] H. Touati. *Performance-oriented Technology Mapping*. PhD thesis, UC Berkeley, Nov. 1990. UCB/ERL M90/109.