Complexity Issues in Gate Duplication

Ankur Srivastava, Ryan Kastner and Majid Sarrafzadeh
Department of Electrical and Computer Engineering
Northwestern University
Evanston, Illinois 60208, USA
ankur,kastner,majid@ece.nwu.edu

Abstract

In this paper we address the complexity issues associated with gate duplication for delay optimization. Gate duplication for general circuits has been proved NP-Complete [1]. In this paper we show that even the local delay optimization by gate duplication is NP-Complete. Local fanout optimization (buffer insertion) for fixed net topology can be solved in polynomial time [9]. Even the global fanout optimization problem has polynomial time complexity if all the pin to pin parameters of a gate are the same and the topology of all the nets is fixed [13]. Hence we show that gate duplication is much harder than buffer insertion.

1 Introduction

Delay optimization is a fundamental goal in logic synthesis. This objective can be achieved in the multiple stages of synthesis flow. This flow could be divided into technology dependent and technology independent phases with the latter preceding the former. In the technology independent stage, the circuit does not have any technology binding whereas technology dependent stage starts after binding (mapping). In this paper we present gate duplication as a strategy for performance optimization. We study the complexity issues in gate duplication and compare it with the complexity of buffer insertion.

Many timing optimization strategies have been proposed over the past few years. Some of the popular strategies for the technology independent phase were suggested in [7, 11] which exploit the concept of restructuring for improvement in circuit performance. Techniques like buffer insertion [4] and gate sizing [14] are popular delay optimization strategies after mapping.

In the past few years the research community has looked at gate duplication extensively as a method of reducing the cut-set of partitions. Strategies of logic duplication for cut-set minimization were suggested in [8] and [10]. In [8] Feduccia and Mattheyses min-cut algorithm has been extended to allow the replication of cells in both the partitions, hence reducing the cut-set of the partition. The strength of gate duplication as a cut-set minimizing strategy has been demonstrated. However applicability of this strategy in reducing the circuit delay has not been studied in detail. One of the few works that we have come across is [6] which addresses the gate duplication problem in a performance driven perspective. It integrates cell replication into a layout driven framework. Benefits of using this strategy on final performance in comparison with conventional transistor sizing techniques have been reported. Another strategy of gate duplication that addresses this problem in the technology independent phase has been proposed in [3].

The global (pertaining to the whole circuit) delay optimization problem by gate duplication is NP-Complete [1]. In this paper we prove the local gate duplication problem to be NP-Complete. The Local Fanout Optimization problem (buffer insertion) with fixed net topology (LFO-NTF) can be solved in polynomial time [9]. Polynomial time solutions exist for the global fanout optimization problem if the delay model does not have separate pin to pin gate parameters and if the net topology is fixed [13]. Hence this problem is in general much harder than buffer insertion.

The rest of this paper is organized as follows. Section 2 deals with the delay model and provides basic definitions. Section 3 reviews the complexity of the global gate duplication problem. Section 4 presents the proof of NP-Completeness for the local gate duplication problem. Section 5 briefly describes some heuristics for solving the gate duplication problem. This is followed by some observations and conclusion in Section 6.

2 Preliminaries

2.1 Delay Models

Given a single output gate g, let $\delta(i, g)$ denote delay from an input pin i of the gate g to the output of g. The load C_g denotes the cumulative capacitance seen at the output of g. It is the sum of the individual input pin capacitances γ_p for all fanouts p of g. Two commonly used delay models for gate level circuits are load dependent delay model **LDDM** and load independent delay model **LIDM** [12]. Figure 1(A) shows the **LIDM**. The delay in the gate g is given by

$$\delta(i, g) = \alpha_{i,g} \tag{1}$$

where $\alpha_{i,g}$ is the intrinsic delay of the gate. Figure 1(B) shows the **LDDM**. The delay through the gate is as follows [13]

$$\delta(i, g) = \alpha_{i,g} + \beta_{i,g} c_g \tag{2}$$

Here,
c_g = load capacitance at the output of the gate g,
$\alpha_{i,g}$ = intrinsic delay from i to output of g,
$\beta_{i,g}$ = drive capability or load coefficient of the path from i to the output of g.

The delay of a path that goes from a primary input(PI) to a primary output(PO) is the sum of the pin-to-pin delays through all the gates lying on the path [13]. The delay in the circuit is the maximum of all the individual path delays.

Let $r(g)$ denote the required time at the output of a gate g. The following equation illustrates the method of computing $r(g)$ if the required times of the fanouts of g are available

$$r(g) = \min_{x \in FO(g)} \{r(x) - \alpha_{g,x} - \beta_{g,x} c_x\} \tag{3}$$

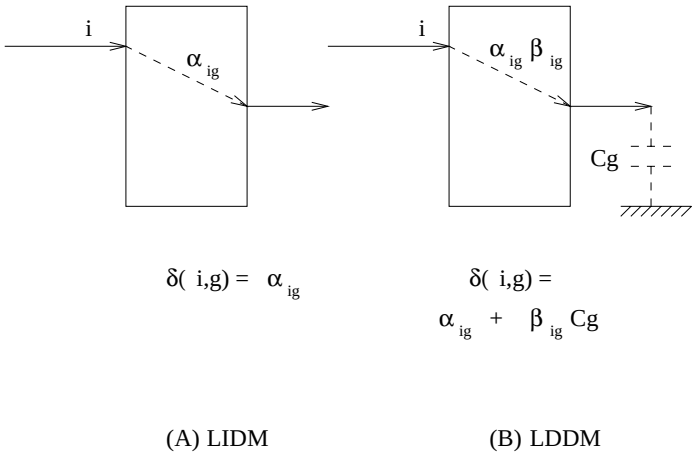

Figure 1: Commonly used delay models

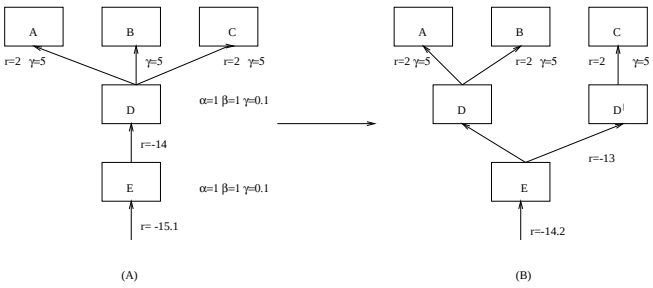

Figure 2: Delay Optimization by Gate Duplication

Here
$\alpha_{g,x}$ = intrinsic delay of gate x w.r.t. the pin connected to g,
$\beta_{g,x}$ = load coefficient of gate x w.r.t the pin connected to g
Required time at the input pin of a gate is define as follows

$$r(i,g) = r(g) - \delta(i,g) \tag{4}$$

In this paper we neglect the wire capacitance. We also assume all gates in the circuit to be single output.

2.2 The Gate Duplication Problem

Gate duplication can be used for delay optimization. The idea is illustrated with the following example. We use the **LDDM** in this example.

Consider the circuit shown in Figure 2(A) in which the parameters α,β and γ along with the required times at the gate outputs have been indicated. The subscript i has been omitted as all the gates have just one input pin. We will show that the delay through this circuit can be improved by duplicating some gates. In the unduplicated case (Figure 2(A)) the capacitive loading C_D = 5+5+5 and C_E = 0.1. Hence the required time at the input of E can be calculated to be -15.1. When D is duplicated (Figure 2(B)), the capacitive loading C_D = 5+5 and C_E = 0.2. Hence the new required time at the input of E becomes -14.2. Gate duplication was hence instrumental in improvement of circuit delay. Duplication caused a reduction in the capacitive loading at D. Although there was an increase in the C_E, the massive decrease in C_D caused the circuit delay to reduce.

It must be noted that in the **LIDM** we cannot formulate the gate duplication problem. This is because the delay model does not have any load dependent parameter. Hence any kind of duplication will not cause the circuit delay to decrease. Henceforth we will be concentrating only on the **LDDM**. In this paper we deal with combinational circuits with single output gates only.

3 Global Optimization by Gate Duplication

The global gate duplication problem is concerned with the entire network.

1. Given a network η consisting of gates and nets
2. Given parameters $\alpha_{i,g}$ $\beta_{i,g}$ and γ_i for each gate g where i is the ith input pin of g
3. Find a duplication strategy that minimizes the circuit delay

This has been shown to be NP-Complete in the **LDDM** [1]. The proof uses Mono3SAT [5] for transformation. This problem remains NP-Complete even when the delay model is simplified to the following

$$\delta_g = \alpha_g + \beta_g c_g \tag{5}$$

The α, β etc parameters are the same for different input pins of a gate g. [13] shows that the global fanout optimization problem (buffer insertion) with fixed net topology (GFO-NTF) can be solved in polynomial time with the above delay model (same pin to pin parameters). Hence global gate duplication is in general harder than global buffer insertion.

Any gate duplication algorithm will have to make at least two decisions on which the final result will depend.

1. Decide the gates to be duplicated
2. Decide the fanouts of the duplicated gates

In the next section we prove that given a duplicated node, the problem of partitioning a set of fanouts between the original and the duplicated node is also NP-Complete.

4 Local Optimization by Gate Duplication

Let us first define the *local gate duplication* problem. Figure 3 illustrates this problem.

1. Given a node n and a set of fanouts that it drives (Figure 3(A))
2. Given the required times at the input pins of all the fanouts
3. Given all gate parameters $\alpha_{i,g}, \beta i, g, \gamma_i$
4. Only gate n can be duplicated
5. Maximize the required time at the input of gate n using gate duplication

In the worst case we will have to look at an exponential number of choices (exponential in the number of fanouts f_n). Here a choice is defined as a partitioning of the fanouts between the gate n and it's duplicate n' (see Figure 3(B)).

In this section we prove that the problem of partitioning a set of fanouts between original and replica for meeting the required time constraint is NP-Complete. The local decision problem (L-GD) can be stated as follows.

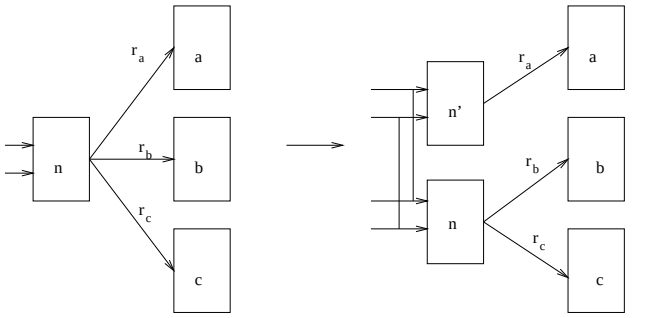

Figure 3: Local gate duplication problem, (A) $r_n = min(r_a, r_b, r_c)$ (B) $r_n = min(r_b, r_c)$, $r_{n'} = r_a$

INSTANCE: Given gates n and n' where n' is a replica of n (with common fanins). Initially n' does not drive any fanouts. Given the FOs of n and the required time at their inputs. Given the $\alpha_{i,n}$ $\beta_{i,n}$ and γ_i for each input pin i of n (or n'), γ for each fanout and a number D.

QUESTION: Does there exist a partitioning of the fanouts FO between n and n' such that the required time at the input pins of n (and n') is at least D.

Proof: Let us first observe that the problem is in NP. Given a partitioning of fanouts, the required times at the input pins of n and n' can be computed in polynomial time. Then we can check if the required constraint is met. Thus L-GD is in NP.
To prove the NP-Completeness, we transform PARTITION [5] to an instance of L-GD.
PARTITION is defined as follows:
INSTANCE: Finite set A, a size $s(a_i) \in Z+$ for each $a_i \in A$ and $\sum_{a_i \in A} s(a_i) = W$

QUESTION: Is there a subset $A' \subseteq A$ such that $\sum_{a_i \in A'} s(a_i)$ = $\sum_{a_i \in A-A'} s(a_i)$ = W/2?

We show that the PARTITION decision is TRUE iff there exists a partitioning of fanouts such that the required time constraint is met.

We transform PARTITION to an instance of L-GD. Figure 4 illustrates the transformation. Let n and n' be the 2 gates. Let the number of fanouts to be partitioned be the number of elements in A. All the gates in the circuit instance have exactly one input pin. Henceforth subscript i will be dropped from all the circuit parameters $\alpha_{i,n}$ $\beta_{i,n}$ and γ_i. The required time at the input pins of all the fanout gates is 0. Parameter γ (input pin capacitance) of the ith fanout is $s(a_i)$. Gates n and n' have $\alpha = 0, \beta = \beta, \gamma = 0$. The required time constraint to be met is $-W\beta/2$. This completes the transformation. It can be seen that this is a polynomial time transformation.

Only If Part: Given a partition A' of A such that

$$\sum_{a_i \in A'} s(a_i) = W/2 \tag{6}$$

We pick each gate $a_i \in A'$ and connect it to n'. All the other gates are connected to n. Capacitive loading of n' is $\sum_{a_i \in A'} s(a_i)$ = W/2. So the required time at the input of n' becomes

$$r_{n'} = -\beta W/2 \tag{7}$$

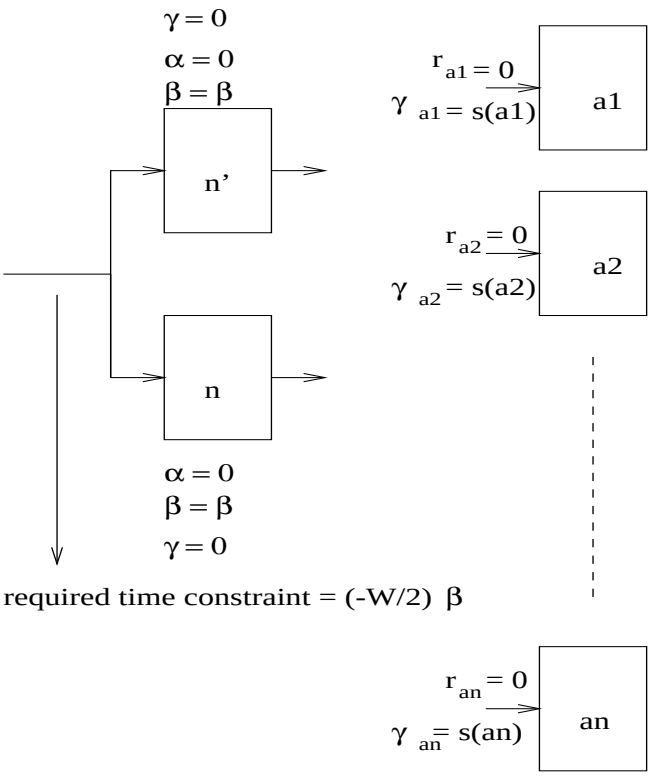

Figure 4: Transformed Circuit Instance

Capacitive loading of n = W - $\sum_{a_i \in A'}$ = W/2. So the required time at the input of n becomes

$$r_n = -\beta W/2 \tag{8}$$

Hence the required time constraint is met.

If Part: Let us assume there exists a partitioning of fanouts such that the required time constraint is met. We prove that this implies the existance of a partition A' of A that satisfies equation 6. First we show that if the required time constraint is met, the capacitive loading at n and n' must be = W/2.

Since the required time constraint is met, the minimum of required times at the input pins of n or n' must be at least $-W\beta/2$. Without loss of generality let us assume that required time at input of n is less than n'. The capacitive loading seen by n must be $\leq$ W/2 (else the constraint will not be met). Since the required time at input of n' is greater than or equal to n, the capacitive loading of n' must be less than or equal to that of n. Hence

$$Capacitive\_loading_n \leq W/2 \tag{9}$$

$$Capacitive\_loading_{n'} \leq Capacitive\_loading_n \leq W/2 \tag{10}$$

Since $Capacitive\_loading_n$ + $Capacitive\_loading_{n'}$ = W, $Capacitive\_loading_n$ = $Capacitive\_loading_{n'}$ = W/2. Hence if the required time constraint is met, the capacitive loading of both n and n' must be W/2. Since the fanout pin capacitances directly corresponds to $s(a_i)$, we can create a partition from the fanouts of n or n' which satisfies equation 6. This completes the proof of NP-Completeness. □

The *local gate duplication* problem defined above will have to first decide if the gate n should be duplicated. If the decision

is in favor of duplication, it will have to partition the fanouts between gates n and n'(it's replica). Since L-GD is NP-Complete, the fanout partitioning problem cannot be solved optimally (in polynomial time). Hence *local gate duplication* is also NP-Complete. Note that local fanout optimization with fixed net topology has polynomial time algorithms.

5 Heuristics for Gate Duplication

In this section we briefly describe a few algorithms for gate duplication that we have implemented. [2] is a generalization of the gate duplication strategy in [3]. It explores the neighborhood of the gate under consideration and evaluates the gain associated with it's duplication. This gain evaluation strategy is more global than the one is [3]. Another strategy based on the structure of dynamic programming has also been discussed in [15]. It traverses the network in topologically sorted order and decides on the gates to be duplicated. These strategies gave improvements as high as 11% over highly optimized results generated by SIS.

6 Conclusion

In the previous section we showed that the problem of partitioning the fanouts between a node and it's replica is NP-Complete. Hence local optimization by gate duplication is also NP-Complete. This is the main contribution of the paper. Close observation reveals that this problem remains NP-Complete even when the delay model is relaxed to the one given by equation 5. This is because the transformation had the same α, β etc parameters for each input pin. Comparing the complexity with buffer insertion

1. Same pin to pin parameters: Global gate Duplication NP-Complete [1], Global buffer insertion is polynomial if net topology is fixed. [13]
2. Same pin to pin parameters: Local gate duplication NP-Complete, Local Fanout Optimization polynomial if net topology is fixed [9]
3. Different pin to pin parameters: Local gate duplication NP-Complete, Local Fanout Optimization polynomial if net topology is fixed [9]

Hence gate duplication is in general much harder than buffer insertion.

References

[1] A. Srivastava and M.Sarrafzadeh. "On the Complexity of Gate Duplication". In *IEEE Trans on Computer Aided Design (Submitted)*.

[2] A. Srivastava, C. Chen and M. Sarrafzadeh. "Timing Driven Gate Duplication in the Technology Independent Phase". In *International Conference on Computer Design (submitted)*, 2000.

[3] C. Chen and C. Tsui. "Timing Optimization of Logic Network using Gate Duplication". In *Proc. Asia and South Pacific Design Automation Conference*, pages 233–236, January 1999.

[4] 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, March 1989.

[5] M.R. Garey and D.S. Johnson. *Computers and Intractability, A guide to the theory of NP Completeness.* W.H. Freeman and Company, New York, 1979.

[6] I. Neumann, D. Stoffel, H. Hartje and W. Kunz. "Cell Replication and Redundancy Elimination During Placement for Cycle Time Optmization". In *Proc. International Conference on Computer Aided Design*, pages 25–30, November 1999.

[7] K.J. Singh, A.R. Wang, R.K. Brayton and A.L. Sangiovanni Vincentelli. "Timing Optimization of Combinational Logic". In *Proc. International Conference on Computer Aided Design*, pages 282–285, November 1988.

[8] C. Kring and A.R.Newton. "A Cell replication Approach to Mincut-Based Circuit Partitioning". In *Proc. International Conference on Computer Aided Design*, pages 2–5, November 1991.

[9] Lukas P.P.P. van Ginneken. "Buffer Placement in Distributed RC-tree Networks for Minimal Elmore Delay". In *Proc of International Symposium on Circuits and Systems*, pages 865–868, December 1990.

[10] M. Enos, S. Hauck and M. Sarrafzadeh. "Evaluation and Optimization of Replication Algorithms for Logic Bipartitioning ". In *IEEE Transactions on Computer Aided Design*, pages 1237–1248, September 1999.

[11] G.De Micheli. "Performance Oriented Synthesis of Large - Scale Domino CMOS Circuits". In *IEEE Transactions on Computer Aided Design*, pages 751–765, September 1987.

[12] R. Murgai. "On the Complexity of Minimum-delay Gate Resizing/Technology Mapping under Load-dependent Delay Model". In *Workshop Handouts, International Workshop on Logic Synthesis*, pages 209–211, June 1999.

[13] R. Murgai. "On the Global Fanout Optimization Problem". In *Proc. International Conference on Computer Aided Design*, pages 511–515, November 1999.

[14] O. Coudert, R. Haddad and S. Manne. "New Algorithms for Gate Sizing: A Comparative Study". In *Proc Design Automation Conference*, June 1996.

[15] A. Srivastava and M. Sarrafzadeh. "Gate Duplication for Performance Optimization". In *Internal Memorandum, Northwestern University*, January 2000.