RUMBLE: An Incremental, Timing-driven, Physical-synthesis Optimization Algorithm

David A. Papa[†♯], Tao Luo[‡], Michael D. Moffitt[♯], C. N. Sze[♯],
Zhuo Li[♯], Gi-Joon Nam[♯], Charles J. Alpert[♯] and Igor L. Markov[†]
[†] University of Michigan, EECS Department, Ann Arbor, MI 48109
[‡] University of Texas at Austin, Department of ECE, Austin, TX 78712
[♯] IBM Austin Research Lab, 11501 Burnet Rd., Austin, TX 78758
iamyou@umich.edu, tluo@ece.utexas.edu {mdmoffitt, csze, lizhuo, gnam, alpert}@us.ibm.com, imarkov@umich.edu

ABSTRACT

Physical synthesis tools are responsible for achieving timing closure. Starting with 130nm designs, multiple cycles are required to cross the chip, making latch placement critical to success. We present a new physical synthesis optimization for latch placement called RUMBLE (Rip Up and Move Boxes with Linear Evaluation) that uses a linear timing model to optimize timing by simultaneously re-placing multiple gates. RUMBLE runs incrementally and in conjunction with static timing analysis to improve the timing for critical paths that have already been optimized by placement, gate sizing, and buffering. Experimental results validate the effectiveness of the approach: our techniques improve slack by 41.3% of cycle time on average for a large commercial ASIC design.

Categories and Subject Descriptors

B.7.2 [**Integrated Circuits**]: Design Aids – *placement and routing*
J.6 [**Computer-Aided Engineering**]: Computer-Aided Design
G.4 [**Mathematical Software**]: Algorithm Design and Analysis

General Terms:

Algorithms, Design, Performance

Keywords:

Timing-driven placement, static timing analysis

1. INTRODUCTION

Physical synthesis is a complex multi-phase process primarily designed to achieve timing closure, though power, area, yield and routability also need to be optimized. Starting with 130nm designs, signals can no longer cross the chip in a single cycle, which means that *pipeline latches* need to be introduced to create multi-cycle paths. This problem becomes more pronounced for 90-, 65- and 45-nanometer nodes, where interconnect delay increasingly dominates gate delay. Hence, the proper placement of pipeline latches is a critical problem for timing closure, especially since there may only be a narrow placement region for the latch (perhaps only a single location) that will close timing.

The place of this optimization in a physical synthesis flow affects the choice of computational techniques. To this end, we review the major phases of such flows following [1, 4].

ISPD'08, April 13–16, 2008, Portland, Oregon, USA.

1. **Global placement.**
2. **Electrical correction** fixes capacitance and slew violations with gate sizing and buffering.
3. **Legalization.** An incremental placement capability that removes overlaps caused by optimization with minimal disturbance to gate locations and timing.
4. **Timing analysis.**
5. **Detailed placement.**
6. **Critical-path optimization.** At this point one can identify most-critical paths and focus on techniques to improve the slack for the worst timing violations. Relevant optimizations include buffering, gate sizing and incremental synthesis [14].
7. **Compression.** When improvements on most-critical paths are no longer possible, one optimizes remaining paths that violate timing constraints. The goal is to *compresses* the timing histogram and reduce number of negative-slack paths that require designer intervention.

The flow can be repeated with net weighting and timing-driven placement to further improve results.

One can think of physical synthesis as progressing with "variable detail / variable accuracy." For example, during global placement, very large changes are made to the design using a coarse objective (such as wirelength) that is oblivious to timing considerations. Later, one may perform more accurate optimization using an Elmore interconnect delay-model with Steiner-tree estimates for net capacitance. As the design begins to converge, one can apply more-expensive, fine-grained buffering along actual detailed routes using a statistical timing model.

Figure 1(a)-(d) illustrates the complications of using global placement to solve the latch placement problem for a single two-pin net. Assume that for all four figures, the source A and sink B are fixed in location and that global placement must find the correct location for the latch. This example is representative of situations in which a fixed block in one corner of the chip must communicate with a block in the opposite corner, and they cannot reach each other in a single cycle. All four placements have equal wirelength, so unless global placement is timing driven, the placement of the latch between A and B is arbitrary. Consider the following scenarios:

- Suppose the placement tool chooses (a), which is the worst location for the latch. In this case, the latch is so far from B that the timing constraint at B cannot be met. This results in a slack on the input net (U) of $+5$ns and a slack on the the output net (V) of -5ns (even after optimal buffering).[1]
- With a second iteration of physical synthesis, timing-driven placement could try to optimize the location of this latch by

[1]The nets in each scenario could include buffers without changing the trends discussed.

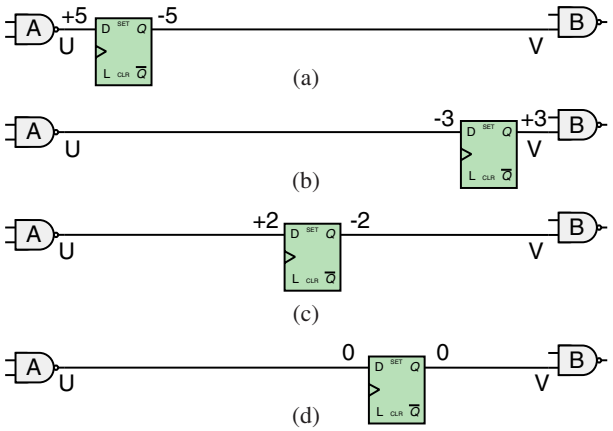

Figure 1: The placement of a pipeline latch impacts the slacks of both input and output paths. A wirelength objective does not capture the timing effects of this situation.

adding net weights. Any net weighting scheme will assign a higher weight to net V than U, resulting in a placement where the latch is very close to B, as in (b). While the timing is improved, there now is a slack violation on the other side of the latch with -3ns of slack on U and $+3$ns on V.

- A global or detailed placer could use a quadratic wirelength objective to handle these kinds of nets, giving the location (c), which, while better than (a) and (b), is still suboptimal.
- To achieve the optimal location with no critical nets (0 slack on U and V), the latch must be placed as shown in (d). In this case, there is only one location that meets both constraints.

This example suggests that wirelength optimization is not well-suited for latch placement, especially when there is little room for error. Instead, one must be able to couple latch placement with timing analysis and model the impact of buffering. In practice, the problem is more complex, and some aspects are not illustrated above. In particular, many latches have buffer trees in the immediate fan-in and fan-out. Such complications pose additional challenges that we address. We make the following contributions.

- We show that a linear-wire-delay model is sufficient to model the impact of buffering for the latch placement problem.
- We develop RUMBLE, a linear-programming-based, timing-driven placement algorithm which includes buffering for slack-optimal placement of individual latches under this model and show its effectiveness experimentally.
- We extend this technique to improve the location of individual logic gates other than latches. Further, we show how to find the optimal location of multiple gates (and latches) *simultaneously*, with additional objectives. Incremental placement of multiple cells requires additional care to preserve timing assumptions, optimizing a set of slacks instead of a single slack, while also biasing the solution towards placement stability. We describe how RUMBLE can handle these situations.
- Our experiments validate the effectiveness of these transforms. We show how these techniques can be used to significantly improve latch placement for a reasonably optimized ASIC design with "do no harm" acceptance criteria that rejects solutions if any quality metrics are degraded. This facilitates the use of RUMBLE later in physical synthesis.

The remainder of the paper is organized as follows. Section 2 discusses background and previous work. Section 3 describes the timing model we use in this work. Section 4 describes how RUMBLE performs timing-driven placement. Section 5 describes the RUMBLE algorithm. Section 6 shows our experimental results. Conclusions are drawn in Section 7.

2. BACKGROUND

The incremental latch placement problem and its multiple movement formulation are explained in Section 4. The high level problem description is: given an optimized design and a small set of gates M (M may consist of a single latch) find new locations for each gate in M and new buffering solutions for nets incident to M such that the timing characteristics of the design are improved.

While moving a cell can improve delay, especially if it has been poorly placed, moving a latch has special significance since it facilitates time-borrowing: reallocating circuit delay from a longer (slow) combinational stage to a shorter (fast) combinational stage. This fact offers a particularly significant boost to our basic approach, and is enhanced even further when surrounding gates are also free to move.

A solution to this problem is called a *transform* using the terminology of [14]. A transform is an optimization designed to incrementally improve the timing. Other examples of transforms include, buffering a single net, resizing a gate, cloning a cell, swapping pins on a gate, etc. The way transforms are invoked in a physical synthesis flow is determined by the *drivers*. For example, a driver designed for critical path optimization may attempt a transform on the 100 most critical cells. A driver designed for compression may attempt at transform on every cell that fails to meet its timing constraints.

A driver has the option of avoiding transforms that may harm the design (e.g., the new buffering solution is worse than the original) and can then reject this solution. This *do no harm* philosophy of optimization has received significant recognition in recent work [5, 11]. The RUMBLE approach adopts this same convention which makes it more trustworthy in a physical synthesis flow.

While no previous work has attempted to solve this particular problem, other works do exist that may be able to help with the placement of poorly placed latches. The authors of [15] propose a linear programming formulation that minimizes downstream delay to choose locations for gates in field-programmable gate arrays (FPGAs). The authors of [6] model static timing analysis (STA) in a linear programming formulation by approximating the quadratic delay of nets with a piecewise-linear function. Their formulation's objective is to maximize the improvement in total negative slack of timing end points. The authors of both approaches conclude that the addition of buffering would improve their techniques [15, 6]. When these transformations are applied at the same point in a physical synthesis flow that we propose, they will be restricted by previous optimizations. When applied somewhat earlier (e.g., following global placement) they are incapable of certain improvements. Namely, downstream optimizations, such as buffer insertion, gate sizing, and detailed placement may invalidate the optimality of latch placement. Therefore, our technique focuses on the bad latch placements that we observed in large commercial ASIC designs after state-of-the-art physical synthesis optimizations. However, we believe that these algorithms are too disruptive to use after routing.

3. THE RUMBLE TIMING MODEL

We now introduce the timing model critical to RUMBLE's success.

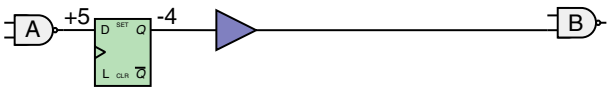

Figure 2: A poorly-placed latch with buffered interconnect. In this case, the buffer must be moved or removed in order to have the freedom to move the latch far enough to fix the path.

Figure 2 shows an intuitive example of the problem when we try to find new locations for movable gates. Similar to Figure 1, the

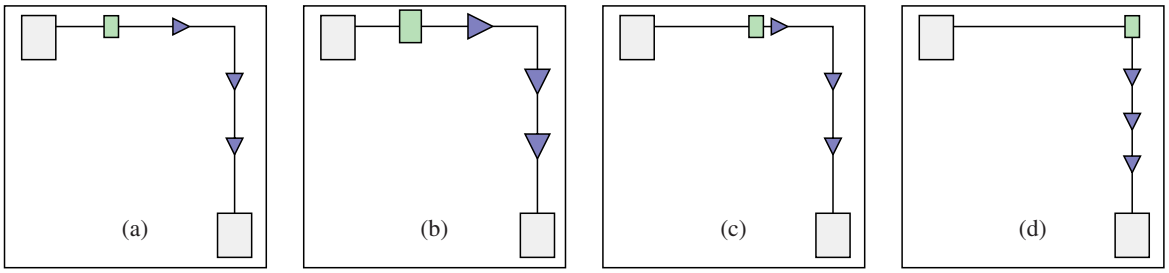

Figure 3: The layout in (a) has a poorly-placed latch, and existing critical path optimizations do not solve the problem. Repowering the gates may improve the timing some in (b), but if it cannot fix the problem, the latch must be moved. Moving the latch up to the next buffer, shown in (c), does not give optimization enough freedom. If we move the latch but do not re-buffer in (d), timing may degrade. Figure 7(d) shows the ideal solution to this problem.

latch has to be moved toward the right for better timing. However, since the latch drives a buffer which is placed next to it, we must move the buffer in order to improve the slack of the latch, and other complications are illustrated by Figure 3. At the same time, the optimal new location of the latch depends on the buffering on the input and output nets. As a result, the optimal approach is to simultaneously move the latch and perform buffering, but it is computationally prohibitive to do so because a typical multiple-objective buffering algorithm runs in exponential time. As mentioned in Section 1, we propose a sequential approach in which we first compute the new locations for a selected set of movable gates based on timing estimation considering buffers. Then, buffering is applied to the input and output nets of the selected movable gates. This approach is practical, effective and efficient and it can be easily integrated into typical VLSI physical synthesis flow. The calculation of optimal movement depends on a simple but effective buffered-interconnect delay model, which is discussed the following section.

3.1 Linear Buffered-Path Delay Estimation

Buffering has become indispensable in timing closure and cannot be ignored during interconnect delay estimation [7, 13, 3]. Therefore, to calculate new locations of movable gates, one must adopt a buffering-aware interconnect delay model that accounts for buffers which are going to be inserted in the future. We found that the linear delay model [10, 3] is best suited in this application. In this model, the delay along an optimally buffered interconnect is

$$delay(L) = L(R_bC + RC_b + \sqrt{2R_bC_bRC}) \quad (1)$$

where L is the length of a 2-pin buffered net, R_b and C_b is the intrinsic resistance and input capacitance of buffers and gates while R and C unit wire resistance and capacitance respectively.

Empirical results in [3] indicate that Equation 1 is accurate up to 0.5% when at least one buffer is inserted along the net. Furthermore, our own empirical results in Section 6.2 suggest a 97% correlation between this linear delay model and the output of an industrial timing analysis tool.

3.2 The Timing Graph

In RUMBLE, a set of movable gates is selected, which must include fixed gates or input/output ports to terminate every path. Fixed gates and I/Os help formulate timing constraints and limit the locations of movables. In Figure 4(a), we assume that new locations have to be computed for the latch and the two OR gates, while all NAND gates are kept fixed.

In the timing graph, each logic gate is represented by a node, while a latch is represented by two nodes because the inputs and outputs of a latch are in different clock cycles and can have different slack values. Each edge represents a driver-sink path along a net and is associated with a delay value which is linearly proportional to the distance between the driver and the sink gate. In other words, we decompose each multi-pin net into a set of 2-pin edges which connect the driver to each sink of the net. This simplification is crucial to our linear delay model and is valid because one of the sinks is usually most critical and all the subtrees off the critical path will be decoupled by buffers. Therefore, the 2-pin edge model in the timing graph can guide the computation of new locations for the movable gates.

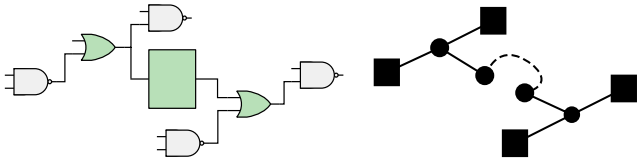

Figure 4: (a) An example subcircuit and (b) corresponding timing graph used in RUMBLE. The AATs or RATs of unmovable objects (squares) are considered known. STA is performed on movable objects (round shapes).

In the timing graph, an edge which represents a timing arc is created only for (1) each connection between the movable gates, and (2) each connection between a movable gate and a fixed gate. This is because we only care about the slack change due to the displacement of movable gates. For the subcircuit in Figure 4(a), the resultant timing graph is shown in Figure 4(b).

For each fixed gate, we assume the required arrival time (RAT) and the arrival time (AT) are fixed. The values of RAT and AT are generated by a static timing analysis (STA) engine using a set of timing assertions created by designers. Please see [9, 12] for an in-depth discussion about STA and the generation of RAT and AT. A movable latch corresponds to two nodes in the timing graph, one for the data input pin and one for the output pin. For the input pin, the RAT is fixed based the clock period. Similarly, the AT is fixed for the latch's output pin. Based on all the fixed RAT and AT at fixed gates and latches, the AT and RAT are propagated along the edges according to the delay of the timing arcs. The values of AT are propagated forward to fan-out edges, adding the edge delay to the AT. On the contrary, RATs are propagated backward to the fan-in edges, subtracting the edge delay from the RAT values. Details of edge delay, RAT and AT calculation will be covered in Section 4.

4. TIMING-DRIVEN PLACEMENT

The goal of RUMBLE is to find new locations for movable gates in some selected subcircuit such that the overall circuit timing improves. Therefore we maximize the minimum slack (i.e., worst slack) of any source to sink timing arc in the subcircuit. We elect this objective in contrast to previous work, because we are targeting critical-path optimization. As such, we prefer 1 unit of worst-

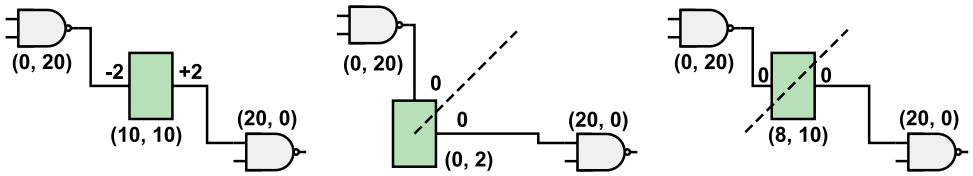

Figure 5: In many subcircuits there are multiple slack-optimal placements. In RUMBLE we add a secondary objective to minimize the displacement from the original placement. This helps to maintain the timing assumptions made initially and reduces legalization issues. (a) shows the initial state of and example subcircuit, (b) a slack-optimal solution commonly returned by LP solvers, all optimal solutions lie on the dotted line and (c) a solution given by RUMBLE that maximizes worst-slack then minimizes displacement.

slack improvement over 2 units of improvement on less-critical nets. Below we introduce the timing-driven placement technique in RUMBLE that directly maximizes minimum-slack. In the following placement formulation we account for the timing impact of our changes by implicitly modeling static timing analysis in our timing graph. In this work, we estimate net length by the half-perimeter wirelength (HPWL) and then scale it to represent net delay. More accurate models are possible.

4.1 Problem Formulation

Consider the problem of maximizing the minimum slack of a given subcircuit G with some movable gates and some fixed gates, or ports.

Let the set of nets in the subcircuit be

$$\mathbf{N} = n_0, n_1, \ldots, n_h \tag{2}$$

Let the set of all gates in the subcircuit (movable and fixed) be

$$\mathbf{G} = g_0, g_1, \ldots, g_f \tag{3}$$

Let the set of movable gates in the subcircuit (a subset of $\mathbf{G}$) be

$$\mathbf{M} = m_0, m_1, \ldots, m_k \tag{4}$$

τ is a technology dependent parameter that is equal to the ratio of the delay of an optimally-buffered, arbitrarily-long wire segment to its length

$$\tau = \frac{delay(wire)}{length(wire)} \tag{5}$$

The following equations govern static timing analysis and are used in the next section. A timing arc is specified for a given net n driven by gate u and having sink v as $n_{u,v}$. The delay of a gate g is D_g.

The Required Arrival Time (RAT) of a combinational gate g is

$$R_g = \min_{o_j : 0 \le j \le m} \{R_{o_j} - \tau * \mathrm{HPWL}(n_{g,o_j}) - D_g\} \tag{6}$$

The Actual Arrival Time (AAT) of a combinational gate g is

$$A_g = \max_{i_j : 0 \le j \le l} \{A_{i_j} + \tau * \mathrm{HPWL}(n_{i_j,g}) + D_g\} \tag{7}$$

For simplicity we assume that the RAT of a latch r is

$$R_r = clock\_period \tag{8}$$

For simplicity we assume that he AAT of a latch r is

$$A_r = 0 \tag{9}$$

The slack of a timing arc $n_{p,q}$ connecting two gates (combinational or sequential, movable or fixed) p and q is

$$S_{n_{p,q}} = R_q - A_p - \tau * \mathrm{HPWL}(n_{p,q}) \tag{10}$$

4.2 The RUMBLE Linear Program

We define a linear-program to maximize the minimum slack S of a subcircuit as follows.

VARIABLES:

$$\begin{array}{llllllll} S & & & & & & & \\ \forall_{m \in M} & : & \beta_x^m & \cup & \forall_{m \in M} & : & \beta_y^m & \cup \\ \forall_{n \in N} & : & U_x^n & \cup & \forall_{n \in N} & : & U_y^n & \cup \\ \forall_{n \in N} & : & L_x^n & \cup & \forall_{n \in N} & : & L_y^n & \cup \\ \forall_{m \in M} & : & R_m & \cup & \forall_{m \in M} & : & A_m & \end{array} \tag{11}$$

Of the Above, β are independent variables for gate locations. The U and L variables represent upper and lower bounds of nets for computing HPWL. R and A compute required and actual arrival times. S is the minimum slack.

OBJECTIVE: **Maximize** S

CONSTRAINTS: For every gate g_j on net n_i

$$U_x^{n_i} \ge \beta_x^{g_j}, \quad U_y^{n_i} \ge \beta_y^{g_j} \tag{12}$$

$$L_x^{n_i} \le \beta_x^{g_j}, \quad L_y^{n_i} \le \beta_y^{g_j} \tag{13}$$

For every movable gate m_i and sink it drives g_j via net n_k

$$R_{m_i} \le R_{g_j} - \tau * (U_x^{n_k} - L_x^{n_k} + U_y^{n_k} - L_y^{n_k}) - D_g \tag{14}$$

For every movable gate m_i and gate that drives it g_j via net n_k

$$A_{m_i} \ge A_{g_j} + \tau * (U_x^{n_k} - L_x^{n_k} + U_y^{n_k} - L_y^{n_k}) + D_g \tag{15}$$

For every timing arc in the subcircuit $n_{p,q}$ on net n_i:

$$S \le R_q - A_p - \tau * (U_x^{n_i} - L_x^{n_i} + U_y^{n_i} - L_y^{n_i}) \tag{16}$$

4.3 Extensions to Minimize Displacement

The linear program of RUMBLE is defined to maximize the minimum slack of a subcircuit. Additional objectives are considered as well, such as total cell displacement, which sums Manhattan distances between cells' original and new locations. We subtract the minimum slack objective from a weighted total cell displacement term to avoid unnecessary cell movement. The weight for the total cell displacement objective, W_d, is set to a small value. Therefore, the weighted total displacement component is used as a tie-breaker and has little impact on worst-slack maximization. Instead, the combined objective is maximized by a slack-optimal solution closest to cells' original locations. During incremental timing-driven placement, minimizing total cell displacement encourages higher placement stability and often translates into fewer legalization difficulties.

Figure 5 shows an example of the RUMBLE formulation with and without the total displacement objectives. The only movable object in Figure 5(a) is the latch. There is an input net n_1 and an output net n_2 connected with the latch. The slack on n_1 is -2 and $+2$ on n_2. Figure 5(b) shows the optimal LP solution without the total displacement objective. The Manhattan net length of n_1 is reduced from 20 to 18 and the net length of n_2 is increased from 20 to 22. Therefore, the new worst slack of the subcircuit was improved from -2 to 0. However, the latch was moved a large distance. In Figure 5(c), including the total displacement objective does not change the optimal slack result. However, the latch displacement is minimized.

We introduce the following variables and constraints to the linear program in order to add the objective to minimize displacement.

DISPLACEMENT VARIABLES:

$$\begin{array}{llllllll} \forall_{m \in M} & : & \delta_x^m & \cup & \forall_{m \in M} & : & \delta_y^m & \cup \\ \forall_{m \in M} & : & \phi_x^m & \cup & \forall_{m \in M} & : & \omega_x^m & \cup \\ \forall_{m \in M} & : & \phi_y^m & \cup & \forall_{m \in M} & : & \omega_y^m & \end{array} \tag{17}$$

DISPLACEMENT CONSTRAINTS:

For every movable gate m_i, $\alpha_x^{m_i}$ and $\alpha_y^{m_i}$ denote the original x- and y-coordinates. The upper and lower bounds of the new and original coordinates ϕ and ω in each dimension are:

$$\begin{array}{llllll} \phi_x^{m_i} & \ge & \beta_x^{m_i}, & \omega_x^{m_i} & \le & \beta_x^{m_i} \\ \phi_y^{m_i} & \ge & \beta_y^{m_i}, & \omega_y^{m_i} & \le & \beta_y^{m_i} \\ \phi_x^{m_i} & \ge & \alpha_x^{m_i}, & \omega_x^{m_i} & \le & \alpha_x^{m_i} \\ \phi_y^{m_i} & \ge & \alpha_y^{m_i}, & \omega_y^{m_i} & \le & \alpha_y^{m_i} \end{array} \tag{18}$$

The displacements δ^{m_i} for a movable gate m_i are defined as

$$\delta_x^{m_i} = \phi_x^{m_i} - \omega_x^{m_i}, \quad \delta_y^{m_i} = \phi_y^{m_i} - \omega_y^{m_i} \tag{19}$$

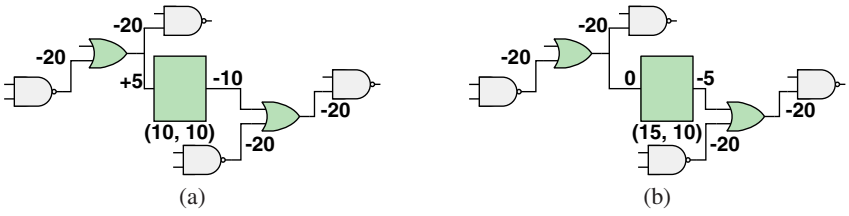

Figure 6: (a) An example subcircuit with an imbalanced latch whose worst-slack cannot be improved. Nevertheless, it is possible to improve timing of the latch while maintaining slack-optimality. By including a FOM component in the objective, the total negative slack can be reduced, as shown in (b).

4.4 Extensions to Improve the Slack Histogram

The minimum slack is the worst slack in a subcircuit. For two subcircuits with identical worst slack, it is possible that one subcircuit has few critical paths with worst slack while the other one has many. A timing optimization has to improve both the worst slack and the overall figure of merit (FOM) in a subcircuit. FOM is defined as the sum of all slacks below a threshold. If the slack threshold is zero, FOM is equivalent to the total negative slack. With the minimum slack as the only objective, to improve a small amount of worst slack may cause a large FOM degradation. Therefore, we must add a FOM component to the optimization objective. The balance between the minimum slack and the FOM is controlled by a parameter W_f, which is set to be relatively small because the worst slack objective is more important.

Figure 6 shows another scenario where the FOM component may help. During the optimization, it may not be always possible to improve the minimum slack of the subcircuit. In that case, we can still reduce the number of critical cells by improving the FOM. In Figure 6, there are three movables in the subcircuit. The minimum slack of the subcircuit is -20, and it is not possible to improve the minimum slack by moving any of the gates. With the additional FOM component in the objective, the FOM of the subcircuit is improved from -90 to -85, as shown in Figure 6(b).

Let S_n denote the slack on net n, the combined objective has the displacement and FOM components
Maximize:

$$\begin{array}{rcl} S & - & W_d \sum_{m \in M} (\delta_x^m + \delta_y^m) \\ & + & W_f \sum_{n:n \in N, S_n < T_s} S_n \end{array} \quad (20)$$

where T_s is the small slack threshold used to compute the FOM.

5. THE RUMBLE ALGORITHM

In this section we discuss the details of the RUMBLE algorithm, which employs the linear program in the previous section to incrementally improve the timing of poorly placed latches.

5.1 Subcircuit Selection

RUMBLE identifies *imbalanced latches*, which we define as those that exhibit positive slack on their inputs and negative slack on their outputs (or vice versa). As illustrated in Figure 1, the movement of any such imbalanced latch has the potential to improve timing, even if all surrounding cells are held fixed. More generally, however, the neighbors and extended neighbors of the targeted latch may also be included to form a set M of movable cells. In our technique, shown in Figure 8, we adopt a basic N-hop neighborhood approach, where any gate within N steps of the imbalanced latch is included in the set of movable cells. This requires both a forward sweep (to collect sinks) and a backward sweep (to collect sources), which are performed in tandem. Those cells that fall $N+1$ steps from the latch form a set P of fixed peripheral nodes.[2]

[2]Variations on this theme, such as metrics that incorporate the degree of neighbors' criticality [15, 8] and the size of the subcircuit bounding box are also possible.

```
(M, P) = BUILD-SUBCIRCUIT-FROM-SEED(Latch L, int N)
M = inputs = outputs = {L}
for i = 1..N + 1
  inputs' = ⋃ ( GET-INPUTS(input) ) ∀input ∈ inputs
  outputs' = ⋃ ( GET-OUTPUTS(output) ) ∀output ∈ outputs
  inputs = inputs', outputs = outputs'
  fixed = output_cone(inputs) ⋂ input_cone(outputs)
  if (i ≤ N) M = M ⋃ inputs ⋃ outputs − fixed
  else P = inputs ⋃ outputs ⋃ fixed   // populates periphery
return (M, P)
```

```
GET-INPUTS(Gate G)
S = ⊘
for each gate G' ∈ pred(G)
  S = S ⋃ TRUE-SOURCE(G')
return S
```

```
GET-OUTPUTS(Gate G)
S = ⊘
for each gate G' ∈ succ(G)
  S = S ⋃ TRUE-SINK(G')
return S
```

```
TRUE-SOURCE(Gate G)
unless (isBuffer(G)) return G
return TRUE-SOURCE(pred(G))
```

```
TRUE-SINK(Gate G)
unless (isBuffer(G)) return G
return TRUE-SINK(succ(G))
```

Figure 8: Subcircuit selection transparently skips buffers when building a neighborhood of movable gates

In contrast to prior work that has assumed operation within a pre-buffering stage, our subcircuit selection algorithm must address the presence of buffers. These buffers will be encountered in our neighborhood selection algorithm, as they are part of the current logic; however, since it is presumed that they will be ripped up when new locations for movables have been determined (a critical assumption that makes our linear-delay model possible), we must prevent their inclusion in our model of the subcircuit. In response, we modify the task of fetching an adjacent gate to transparently skip these buffers, omitting them from the set M. The recursive functions TRUE-SOURCE() and TRUE-SINK() in Figure 8 provide this additional level of indirection, returning only those combinational gates that reflect the logical structure of the subcircuit.

As noted in [5], the process of extracting gates to form a subcircuit suffers from complications when subpaths of combinatorial logic between peripheral nodes are not modeled. These subpaths may introduce additional timing constraints that, if left absent from the model, could invalidate the optimality of the solution. Hence, we intersect the transitive cones of logic between inputs and outputs to capture these paths, obtaining a so-called *convex* subcircuit. To improve runtime, we limit the depth of these cones to a reasonably small constant, as opposed to the exhaustive expansion in [5].

5.2 The "Do no harm" Philosophy

After gates are moved it is likely that timing has degraded due to, for example, a capacitance violation on a long wire. The subcircuit must be examined and its interconnect improved through physical synthesis optimizations, which might include resizing gates and inserting buffers for delay or electrical considerations on nets.

Even though the linear program of Section 4.2 can be solved op-

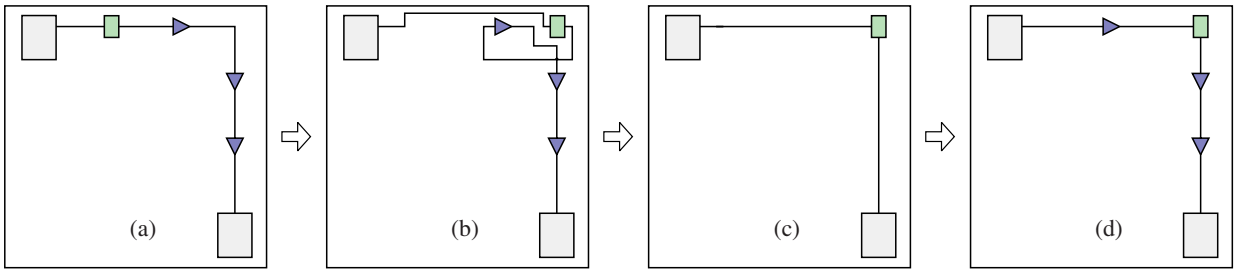

Figure 7: The RUMBLE algorithm proceeds by (a) selecting a subcircuit to work on. An LP is formulated and solved, with movable gates being relocated as shown in (b). Existing repeater trees are no longer appropriate, and are subsequently removed in (c). Finally, the nets are re-buffered, forming the final subcircuit shown in (d).

timally, it does not account for all the complexities of interconnect optimization. The linear program is an abstraction of the subcircuit timing that models the physical synthesis optimizations (e.g., virtual-buffering) by setting a wire delay constant that reflects an estimate of what timing will be after physical synthesis optimizations are performed. Despite the high correlation to more accurate timing models in experimental results, the RUMBLE model could turn out to be too optimistic and its solution might result in a timing degradation. For example, nets can cross congested regions or blockages where no nearby legal locations for buffers can be found. As a result, legalization could create a timing degradation.

When running RUMBLE in our physical synthesis flow, we mitigate the harmful effects of legalization by finding legal locations for gates and buffers when moving or inserting them. Insisting on legal locations can also contribute to a degradation not anticipated by the RUMBLE model. Fortunately, RUMBLE can examine the timing implications of its changes before committing to them. It simply stores the initial state of the subcircuit, and restores it if a timing degradation occurs. In this way, RUMBLE will "do no harm" to the circuit by ensuring that whatever solution it keeps is no worse than what existed before.

5.3 The RUMBLE Algorithm

Figure 9 shows pseudocode for the RUMBLE algorithm. It works on a set of movable gates given as input. First, the subcircuit that is necessary for incremental placement is extracted (for a single movable, it is the one-hop neighborhood of the input gates). During this process, buffers are passed-over (ignored) as described in Section 5.1. Next, RUMBLE takes a snapshot of the timing which is used to measure improvement later. Lines 3 and 4 store the state of the circuit (gates and nets) in preparation for a possible undo of the optimizations we are considering. Once the initial state is safely stored, lines 5-7 use the linear program of Section 4 to compute new gate locations, followed by buffer removal. If the model shows improvement we continue, and all physical synthesis optimizations, including buffering, are lumped into a function call in Line 9. Lines 10-13 measure improvement, and in the case of timing degradation, undo all changes.

6. EXPERIMENTAL RESULTS

RUMBLE is implemented in C++ (compiled with GCC 4.1.0) and integrated into an industrial physical synthesis flow. For our experiments, we examined an already optimized 130nm commercial ASIC with clock period 2.2ns and 3 million objects. We first examined the most critical latches and then filtered out the ones where the latch was already well placed. We use the algorithm from [2] to perform buffering after the cells have been moved. In practice, the LP-solving technique from RUMBLE requires only 17 milliseconds; the buffering algorithm dominates the runtime (over 75%). Since the overall runtime is dependent on the choice of the buffering algorithm we omit the (trivial) runtimes from our tables. Note that the "do no harm" approach of Section 5.2 is applied to all experiments, preventing timing degradation in our tables.

```
RUMBLE (Gate movable)
1    subcircuit = Build-Subcircuit-From-Seed(movable, 0)
2    before_timing = measure_timing(subcircuit)
3    initial_solution.create_interconnect_cache(subcircuit)
4    initial_solution.before_locs = get_locations(movables)
5    Build LP the RUMBLE linear program for subcircuit
6    after_locs = LP.solve()
7    set_gates_locations(movables, after_locs)
8    initial_solution.rip_up_buffers()
9    phys_syn_opt(movables, initial_solution.get_nets());
10   after_timing = measure_timing(subcircuit)
11   if(after_timing worse than before_timing)
12     set_locations(movables, initial_solution.before_locs)
13     initial_solution.restore_interconnect()
```

Figure 9: The RUMBLE algorithm for moving one latch.

6.1 Re-buffering in RUMBLE

Previously published LP techniques for timing-driven placement do not allow for re-buffering during optimization. Instead, they are either applied at a stage in the physical synthesis flow before buffers have been inserted, or they do not differentiate the buffers from other gates. This first experiment is designed to show how important it is to rip up buffers before replacing gates then re-buffer.

We modified our pseudocode in Figure 8 so that the $isBuffer()$ function always returns false. The effect of this is to stop "seeing through" the buffers, and instead to consider them fixed timing endpoints. This setup results in something similar to the work of [15]. We then calculate a new location for each latch with the LP in Section 4. The final change is to skip line 9 of Figure 9, i.e., do not re-buffer. We call this algorithm KEEP-BUFFERS.

Table 1 shows the results of RUMBLE on a single latch compared with KEEP-BUFFERS. Column 1 shows the name of the benchmark and columns 2 and 5 show worst-slacks in picoseconds before optimization. Columns 3 and 6 show the slacks after optimization of KEEP-BUFFERS and RUMBLE respectively. Columns 4 and 7 show the improvements of each technique.

From the table we observe the following:

- Despite not ripping up buffers, KEEP-BUFFERS is still able to improve solution quality for nine out of ten testcases, though the improvement is never more than 220ps.
- When buffer rip-up and re-buffering is allowed, RUMBLE is able to significantly outperform KEEP-BUFFERS for all ten testcases. On average the improvement is 7.4x greater.
- While KEEP-BUFFERS improves slack by an average of 123ps, RUMBLE improves slack by 908ps, which validates how important it is to rip-up buffers so that they do not anchor the latch into a artificially small region.

Implications of keeping buffers						
	KEEP-BUFFERS Slack (ps)			RUMBLE Slack (ps)		
Subcircuit	orig	new	imprv.	orig	new	imprv.
latch A0	-1480	-1318	**162**	-1480	26	**1506**
latch A1	-1268	-1066	**202**	-1268	186	**1454**
latch A2	-1020	-939	**80**	-1020	-791	**229**
latch A3	-953	-766	**187**	-953	-390	**563**
latch A4	-897	-677	**220**	-897	356	**1253**
latch A5	-848	-746	**101**	-848	-278	**570**
latch A6	-690	-690	**0**	-690	395	**1085**
latch A7	-645	-586	**59**	-645	-19	**626**
latch A8	-633	-560	**74**	-633	290	**923**
latch A9	-610	-466	**144**	-610	262	**872**
avg	*-904*	*-782*	*123*	*-904*	*4*	*908*

Table 1: Keeping buffers instead of removing and reinserting them degrades RUMBLE's performance.

6.2 Accuracy of the RUMBLE Timing Model

Theoretical results published by Otten [10] and discussed in Section 3 indicate that optimal buffer insertion on a 2-pin net results in a wire delay that is linearly-proportional to its length. The RUMBLE model heavily relies on these results.

Table 2 compares the model-predicted values for subcircuit slack to values measured by running a commercial static timing analyzer. Measurements are taken after the RUMBLE LP is solved, the latches are moved and connected nets are buffered. Columns 2-4 report the initial, final, and improvement in worst-slack of the subcircuit measured by the timing model presented in Section 3. Columns 5-7 report the same metrics measured by the STA engine.

Model timing vs. reference timing						
	Model slack (ps)			Subcircuit slack (ps)		
Subcircuit	orig	new	imprv.	orig	new	imprv.
latch A0	-1799	-48	**1751**	-1480	26	**1506**
latch A1	-1509	65	**1574**	-1268	186	**1454**
latch A2	-1113	-868	**245**	-1020	-791	**229**
latch A3	-1147	-527	**620**	-953	-390	**563**
latch A4	-1090	180	**1269**	-897	356	**1253**
latch A5	-945	-295	**650**	-848	-278	**570**
latch A6	-920	320	**1241**	-690	395	**1085**
latch A7	-886	49	**935**	-645	-19	**626**
latch A8	-913	213	**1126**	-633	290	**923**
latch A9	-800	397	**1198**	-610	262	**872**
avg	*-1112*	*-51*	*1061*	*-904*	*4*	*908*

Table 2: The RUMBLE model accurately predicts the solution quality improvements in the reference timing model.

We make the following observations:

- On average, the RUMBLE model overestimates the actual timing improvement by about 15%. This makes sense since it assumes an optimal ideal buffering will be achievable, but this is not always the case, especially for multi-sink nets.
- However, if one compares actual improvement to model improvement, there is a 97% correlation, suggesting that the model is reasonable enough to justify the latch location.

We now show how RUMBLE actually improves the design's timing characteristics.

6.3 RUMBLE on a Single Latch

Given that we are solving a new physical synthesis problem, existing solutions are scarce. Therefore, we first consider straightforward approaches to solve this problem. One possibility is to take the *center-of-gravity* (COG) of adjacent pins. A timing-driven improvement of the center-of-gravity technique weights each pin by its slack. A reasonable version of this heuristic works in the following way. For a slack threshold T_s (see Section 4.4), let the weight w of a pin p with slack S_p be:

$$w_p = \begin{cases} 1 + |S_p - T_s| & S_p < 0 \\ \max(0.1, 1 - |S_p - T_s|) & S_p \geq 0 \end{cases}$$

Then compute the x-coordinate of movable gate m as the weighted average of the x-coordinates of the set of neighboring pins P.

$$m_x = \frac{\sum_{p \in P} w_p p_x}{\sum_{p \in P} w_p}$$

and similarly for the y-coordinate.

We implemented the above COG technique within the RUMBLE framework in place of the LP solver presented in Section 4. We still allow COG the benefits of ripping up buffers, and reinserting them after the latches are moved. Table 3 shows a comparison between RUMBLE and slack-weighted COG on 10 latches. Column 1 shows the same latches as reported in Table 2. Columns 2-4 show the initial and final slacks, and improvement for COG. Columns 5-7 show the same for RUMBLE.

Center-of-gravity vs. RUMBLE						
	COG Slack (ps)			RUMBLE Slack (ps)		
Subcircuit	orig	new	imprv.	orig	new	imprv.
latch A0	-1480	-527	**953**	-1480	26	**1506**
latch A1	-1268	-203	**1065**	-1268	186	**1454**
latch A2	-1020	-800	**219**	-1020	-791	**229**
latch A3	-953	-615	**338**	-953	-390	**563**
latch A4	-897	-78	**819**	-897	356	**1253**
latch A5	-848	-319	**529**	-848	-278	**570**
latch A6	-690	-690	**0**	-690	395	**1085**
latch A7	-645	-645	**0**	-645	-19	**626**
latch A8	-633	-633	**0**	-633	290	**923**
latch A9	-610	67	**677**	-610	262	**872**
avg	*-904*	*-444*	*460*	*-904*	*4*	*908*

Table 3: Comparison of RUMBLE's LP to a slack-weighted center-of-gravity technique.

We observe the following:

- For all ten cases, RUMBLE generates a better solution than COG. For three of the cases, COG could not improve the latch placement. These new solutions are rejected by the driver so as not to make the design worse.
- On average, COG improves slack by 20.9% of the 2.2ns cycle time, whereas RUMBLE improves slack by 41.3%. This shows that one must incorporate slack constraints on cells incident on the latch to achieve the most balanced solution.

6.4 Optimizing Multiple Gates Simultaneously

For our final experiment, we show how an even better solution can be obtained when one allows cells close to the latch to move. We show the effectiveness of this technique on two sets of circuits.

- **One-hop** subcircuits include every gate (while ignoring buffers and inverters) incident to the latch of interest that shares an incident net with the latch. Typically this results in 4 or 5 gates being moved.
- **Two-hop** subcircuits in addition include all non-buffer and inverter cells incident to cells in the one-hop neighborhood.

We compare this technique to iterated single-move RUMBLE, where we pick each cell in the neighborhood and solve the LP for that particular cell, fix it, and then move to the next cell. The experiment is designed to show that multiple cells need to be optimized simultaneously to obtain the best results.

To measure the improvement one must now consider the slacks of all cells that may be moved, and the objective becomes to improve the worst slack of the entire subcircuit. However, when one cannot improve the most critical path, the other paths may have

Iterated RUMBLE vs. RUMBLE: 1-hop												
	Iterated single-move RUMBLE						Multi-move RUMBLE					
	Slack (ps)			FOM (ps)			Slack (ps)			FOM (ps)		
Subcircuit	orig	new	imprv.	orig	new	imprv.	orig	new	imprv.	orig	new	imprv.
subcircuit B0	-1542	-1542	**0**	-6091	-6091	**0**	-1542	-130	**1412**	-6091	-130	**5962**
subcircuit B1	-1501	-277	**1223**	-5924	-277	**5647**	-1501	55	**1556**	-5924	0	**5924**
subcircuit B2	-1240	-1240	**0**	-4354	-4354	**0**	-1240	-980	**261**	-4354	-4044	**310**
subcircuit B3	-848	-278	**569**	-2523	-812	**1710**	-848	-279	**569**	-2523	-813	**1709**
subcircuit B4	-690	-79	**612**	-4090	-79	**4011**	-690	202	**893**	-4090	0	**4090**
subcircuit B5	-690	48	**739**	-2053	0	**2053**	-690	290	**980**	-2053	0	**2053**
subcircuit B6	-645	-18	**627**	-1921	-32	**1889**	-645	301	**945**	-1921	0	**1921**
subcircuit B7	-595	86	**681**	-1937	0	**1937**	-595	503	**1098**	-1937	0	**1937**
subcircuit B8	-444	-444	**0**	-889	-889	**0**	-444	-92	**352**	-889	-191	**698**
subcircuit B9	-418	-46	**372**	-857	-46	**811**	-418	6	**424**	-857	0	**857**
avg	*-861*	*-379*	*482*	*-3064*	*-1258*	*1806*	*-861*	*-12*	*849*	*-3064*	*-518*	*2546*

Table 4: RUMBLE simultaneously moving a *one-hop* neighborhood compared to iteratively moving the same gates individually.

Iterated RUMBLE vs. RUMBLE: 2-hop												
	Iterated single-move RUMBLE						Multi-move RUMBLE					
	Slack (ps)			FOM (ps)			Slack (ps)			FOM (ps)		
Subcircuit	orig	new	imprv.	orig	new	imprv.	orig	new	imprv.	orig	new	imprv.
subcircuit C0	-719	-719	**0**	-8313	-8313	**0**	-719	-675	**44**	-8313	-5028	**3285**
subcircuit C1	-719	-719	**0**	-8004	-8004	**0**	-719	-653	**66**	-8004	-4386	**3617**
subcircuit C2	-690	-79	**612**	-4090	-79	**4011**	-690	314	**1004**	-4090	0	**4090**
subcircuit C3	-690	-79	**612**	-4090	-79	**4011**	-690	337	**1027**	-4090	0	**4090**
subcircuit C4	-681	-349	**333**	-3865	-349	**3516**	-681	-158	**524**	-3865	-158	**3707**
subcircuit C5	-645	-91	**554**	-3767	-306	**3462**	-645	371	**1015**	-3767	0	**3767**
subcircuit C6	-645	-33	**612**	-3767	-52	**3716**	-645	324	**969**	-3767	0	**3767**
subcircuit C7	-318	-318	**0**	-940	-940	**0**	-318	531	**848**	-940	0	**940**
subcircuit C8	-490	227	**716**	-966	0	**966**	-490	466	**956**	-966	0	**966**
subcircuit C9	-217	-217	**0**	-652	-652	**0**	-217	60	**277**	-652	0	**652**
avg	*-581*	*-238*	*344*	*-3846*	*-1877*	*1968*	*-581*	*92*	*673*	*-3846*	*-957*	*2888*

Table 5: RUMBLE simultaneously moving a *two-hop* neighborhood compared to iteratively moving the same gates individually.

room for improvement. We use FOM to measure the total improvement of all the slacks in the subcircuit.

Tables 4 and 5 compare iterating RUMBLE over each gate one at a time versus RUMBLE moving multiple gates simultaneously. Columns 2-4 show the original and final slack, and the slack improvement for iterated single-move RUMBLE, while columns 5-7 show the corresponding FOM measurements for a zero-slack threshold. Columns 8-13 show the same measurements for multi-move RUMBLE. We make the following observations:

- Multi-move RUMBLE is clearly more effective than iterative RUMBLE both for one- and two-hop neighborhoods. In fact, for six out of ten one-hop subcircuits and for seven out of ten two-hop circuits, multi-move actually brought the FOM down to zero, meaning it fixed all the timing violations. Iterative single move was able to fix two and four respectively.
- On average, the worst-slack improvements were 849ps and 673ps respectively for one- and two-hop subcircuits. The diminished improvement for larger subcircuits is likely because we are including more nets, some of which cannot be improved as much as those connected to the imbalanced latch (Figure 6 has an example).
- Solving the LP takes 53ms for one-hop and 325ms for two-hop, on average.

7. CONCLUSIONS

In this work we observe that wirelength-driven placement leads to particularly poor timing of "pipeline latches" in modern physical design flows. To address this challenge, we developed RUMBLE — a linear-programming based, incremental physical synthesis algorithm that incorporates timing-driven placement and buffering. The latter justifies RUMBLE's linear-delay model which exhibited a 97% correlation to the reference timing model in our experiments. Empirically this delay model is accurate enough to guide optimization; RUMBLE improves slack by 41.3% of cycle time on average for a large commercial ASIC design.

The LP used in RUMBLE is general enough to optimize multiple gates and latches simultaneously. However, when moving multiple gates considering only the slack objective, we encountered two challenges: placement stability and FOM degradations. We present our extensions to address these problems directly in our LP objective. With these additions, moving several gates simultaneously improves upon RUMBLE used iteratively on the same movables.

8. REFERENCES

[1] C. J. Alpert, C. Chu, and P. G. Villarrubia, "The Coming of Age of Physical Synthesis," *ICCAD*, 2007, pp. 246-249.

[2] C. J. Alpert et al., "Fast and Flexible Buffer Trees that Navigate the Physical Layout Environment," *DAC*, 2004, pp. 24-29.

[3] C. J. Alpert et al., "Accurate Estimation of Global Buffer Delay Within a Floorplan," *TCAD* 25(6), 2006, pp. 1140-1146.

[4] C. J. Alpert, et al., "Techniques for Fast Physical Synthesis," *Proc. IEEE* 95(3), 2007, pp. 573-599.

[5] K-H. Chang, I. L. Markov and V. Bertacco, "Safe Delay Optimization for Physical Synthesis," *ASPDAC*, 2007, pp. 628-633.

[6] A. Chowdhary et al., "How Accurately Can We Model Timing In A Placement Engine?," *DAC*, 2005, pp. 801-806.

[7] J. Cong, L. He, C.-K. Koh and P. H. Madden, "Performance Optimization of VLSI Interconnect Layout," *Integration: the VLSI Journal*, 1996, vol. 21, pp. 1-94.

[8] T. Luo, D. Newmark and D. Z. Pan, "A New LP Based Incremental Timing Driven Placement for High Performance Designs," *DAC*, 2006, pp. 1115-1120.

[9] R. Nair, C. Berman, P. Hauge and E. Yoffa, "Generation of Performance Constraints for Layout," *TCAD* 8(8), 1989, pp. 860-874.

[10] R. Otten, "Global Wires Harmful?," *ISPD*, 1998, pp. 104-109.

[11] H. Ren et al, "Hippocrates: First-Do-No-Harm Detailed Placement" *ASPDAC*, 2007, pp. 141-146.

[12] S. Sapatnekar, "Timing," Springer-Verlag, New York, 2004.

[13] P. Saxena, N. Menezes, P. Cocchini and D. A. Kirkpatrick, "Repeater Scaling and Its Impact on CAD," *TCAD* 23(4), 2004, pp. 451-463.

[14] L. Trevillyan et al., "An Integrated Environment for Technology Closure of Deep-submicron IC Designs," *IEEE Des. Test Comput.*, 2004, vol. 21, no. 1, pp. 14-22.

[15] Q. Wang, J. Lillis and S. Sanyal, "An LP-Based Methodology for Improved Timing-Driven Placement," *ASPDAC*, 2005, pp. 1139-1143.