Glitch Power Minimization by Gate Freezing

L. Benini*  G. De Micheli #  A. Macii@  E. Macii@  M. Poncino*  R. Scarsi*

1 Politecnico di Torino
Torino, Italy 10129  # Stanford University
Stanford, CA 94305  @ Università di Bologna
Bologna, Italy 40136

Abstract

This paper presents a technique for glitch power minimization in combinational circuits. The total number of glitches is reduced by replacing some existing gates with functionally equivalent ones (called F-Gates) that can be "frozen" by asserting a control signal. A frozen gate cannot propagate glitches to its output. An important feature of the proposed method is that it can be applied in-place directly to layout-level descriptions; therefore, it guarantees very predictable results and minimizes the impact of the transformation on circuit size and speed.

1 Introduction

Minimizing glitches at the gate level is a complex task, because it is difficult to estimate the impact that the transformations can have on glitch power. Two different approaches have been taken to solve this problem. Networks can be designed using a glitch-free implementation style (e.g., Shannon [1] circuits); alternatively, a non-glitch-free implementation can be optimized to reduce the number of glitches. The first solution has a major limitation: glitches are eliminated, but the constraints imposed by the design style may lead to circuits that are less power efficient than those realized with standard static CMOS gates. The second approach is hindered by the uncertainties in the estimation of the cost function that drives the optimization procedure. In other words, it is difficult to estimate if the changes in the network introduced to minimize glitching are useful or not, because they may modify the delay distribution of the circuit in an unpredictable fashion.

We follow the second avenue of attack, but we overcome the predictability problem by posing tight constraints on the amount of network perturbation that can be introduced by the glitch minimization procedure. More in detail, we propose an optimization method that operates in-place on a layout-level description. We perform minor modifications of the netlist that can be implemented on the placed and routed circuit only by applying partial re-wiring of a few signal nets. The cost function that drives the optimization is very accurate, because glitch and total power estimation are carried out on a network with accurate back-annotation of wiring loads.

Automatic techniques for glitch minimization in logic circuits have been the subject of intensive research in the past. Some of them have been developed for specific applications [4], while others are general-purpose [5, 6]. In addition, guarded evaluation [7], although not explicitly targeting glitch power minimization, has some affinity with the gate freezing idea we present in this paper. Differently from gate freezing, the techniques above are applied before placement and routing.

We have applied the gate freezing procedure to the Iscas'85 benchmarks [2], as well as to some examples taken from the Msc's91 suite [3]. We have achieved an average glitch power reduction of 14.0%, resulting in an average total power savings of 6.3%. Area and speed of the circuits are virtually unchanged. Obviously, gate freezing only targets circuits with high glitching; therefore, its applicability is of limited interest in the cases where spurious transitions have a negligible impact on the global power balance.

2 Gate Freezing

The flow of the gate freezing procedure starts from a placed and routed combinational circuit. We assume that wire loads have been extracted and back-annotated into the gate-level model of the circuit. Accurate switch-level simulation of user-specified patterns is performed to obtain a detailed statistic of the glitching activity. Static gate-level timing analysis is also performed, and arrival and required times are computed for each node in the network. The information on glitching and on timing constraints is subsequently exploited in the gate-selection step. A fraction of the gates in the netlist is selected. We choose non-critical gates with high glitching and high fanout. The selected gates are then replaced with functionally equivalent F-Gates with one additional control input. Whenever the control signal is low, the gate is not sensitive to input transitions. Glitches on the output of the gate can thus be eliminated by first keeping the control signal low at the beginning of the clock cycle until the input signals of the gate have stabilized to their final value, and then by raising it and keeping it high until the end of the clock cycle. The cost of the generation of the control signals is amortized by clustering the F-Gates. Gates in a cluster share a single control signal. The last step in the gate freezing procedure is the incremental modification of the layout. The selected gates are replaced by F-Gates, the drivers for the control signals are instantiated, and the control signals are routed.

A practical implementation of gate freezing, to be successful, must satisfy two critical requirements. First, the overhead for the generation of the control signals should be more than paid off by the power saved with gate freezing. Second, the physical implementation of the circuit should be minimally modified by the transformation. Hence, we need low-overhead implementations for F-Gates and C-Signals generation circuitry; moreover, we need effective algorithms for selecting target gates and for grouping them into clusters that share a common C-Signal. The next two subsections are dedicated to the description of efficient implementations for F-Gates and C-Signals. The algorithms for gate selection and clustering are described in Section 3.
2.1 F-Gates

Once a target cell in the tech-mapped circuit has been selected, it is replaced by a modified library cell (the F-Gate) whose output can be selectively "frozen" with the purpose of reducing the amount of glitching.

The basic modification of a generic CMOS library cell is shown in Figure 1. It consists of the insertion of a n-type transistor in series with the n-network. The gate input of this n-type transistor is driven by the control input C.

![Figure 1: Basic Transformation of a Library Cell.](image)

The behavior of the modified gate is quite intuitive: When the control input C is high, the gate operates normally; on the other hand, when C is low, the gate output is disconnected from the ground and, therefore, it can never be discharged to the logic value 0.

In this configuration, the output of the gate is only partially "frozen"; in fact, only the 1 to 0 transition is actually forbidden, whereas the gate output can still exhibit the 0 to 1 transition. This may occur for any input configuration that is supposed to force a 1 on the output. In other terms, a low control input C will never allow a gate output that is at the logic value 1 to make a transition.

The above considerations imply that we do not guarantee to completely filter out transitions on the gate output. One trivial way to obtain a total filtering would be to mimic the structure of dynamic gates, and symmetrically insert a p-type transistor before the supply connection, in series with the p-network; in this case, the gate terminal of this p-transistor should be controlled by the other phase, C, of the control input C.

This solution would clearly suppress any transition on the gate output, and solve the above limitation of the single n-transistor solution. However, there are some considerations that make the introduction of the p-transistor undesirable:

i) The load on the signal C is doubled for any gate such signal must drive; ii) Both phases of signal C are required; iii) The area of the modified gate is sensibly larger than in the case of the single n-transistor; this is because, to guarantee the proper output signal levels, the p-transistor must be larger than the n-transistor.

The single n-transistor solution is less intrusive with respect to the size of the library cell and its speed; therefore, it is preferable to the complementary one. Nevertheless, the single n-transistor configuration will obviously be slightly larger and slower than the original cell.

2.2 C-Signal Generation

As described at the beginning of this section, static timing analysis is carried out on the initial mapped circuit in a pre-processing phase. For each gate g, arrival time \( A_T(g) \), required time \( RT(g) \), and slack \( S(g) \) are available.

Once the nodes in the network have been ranked according to their suitability using the criteria described in Section 3.1, we must restrict the choice for the replacement with a modified cell only to gates that are not on the critical path (i.e., gates with non-null slack). This is because, even though the modified cells are only slightly slower than the standard ones, replacing a critical gate will cause the cycle time of the circuit to increase.

We now derive the timing and functional conditions that define the behavior of the control signal C for a candidate gate g. In the following, let T denote the clock period, g the candidate gate, and \( A_T(g) \) its arrival time. Moreover, let \( X = \{x_1, \ldots, x_n\} \) denote the inputs of g, and \( A_T(x_i), RT(x_i), i = 1, \ldots, n \) their corresponding arrival and required times.

We should observe first that all transitions occurring prior to the arrival time are glitches, and can thus be suppressed. This implies that C, the control signal of the modified gate that will replace g, can be held at 0 in the time interval between the beginning of the clock cycle and the time \( \Delta \) that equals the latest arrival time of the inputs of g. In symbols:

\[
\Delta(g) = \max_{x_i} A_T(x_i) \tag{1}
\]

At time \( \Delta(g) \), gate g will have all of its inputs "ready", and will be ready to propagate its final value; this implies that signal C should go high at time \( \Delta(g) \) in order to allow gate g to exhibit a correct temporal behavior.

However, the control signal C does not actually need to go high exactly at \( \Delta(g) \). As a matter of fact, if g is non-critical, all of its inputs will be non-critical as well. Therefore, all the inputs of g will have some slack. What we must then guarantee is that C goes to 1 (i.e., g is free to evaluate) before the time, which is the earliest required time of all inputs of g. In symbols:

\[
\tau(g) = \min_{x_i} RT(x_i) \tag{2}
\]

This allows a safety margin for the transition of the control signal, because we have a don't care region between \( \Delta(g) \) and \( \tau(g) \), where we can decide whether to raise signal C or not.

In summary, signal C should have a timing behavior as the one depicted in Figure 2. C can be thought of as a delayed copy of the clock, yet with a different duty cycle \( D_C = \frac{\Delta(g)}{T} \). In the picture, the shaded area shows the slack for the 0 to 1 transition on C. With such signal, we can guarantee that all the glitches before the latest arrival time of the gate's inputs are filtered out; the only spurious transition that can propagate through a gate is the 0 to 1 transition mentioned in Section 2.1. The shaded area in the figure shows this glitch-free portion of a clock cycle.
Signal $C$ can be derived by proper filtering of the clock signal, $ck$. As mentioned above, in the design of the circuit for the generation of the control signal, we limit ourselves to considering the $\Delta$'s, and use the slack of the control signal, $-\Delta$ only as a safety margin.

We can derive the following relations between the clock signal $ck$ and the control signal $C$. If the time $\Delta$ is smaller than the fraction of the clock period with $ck = 1$, $C$ can be obtained as: $C = ck' + ck_\Delta$, where $ck'$ is the inverse of the clock signal, and $ck_\Delta$ is the clock signal delayed by an amount of $\Delta$. Conversely, if $\Delta$ is greater than the fraction of the clock period with $ck = 1$, $C$ is computed as: $C = ck' - ck_\Delta$. The above relations are shown by means of timing diagrams in Figure 3-a and 3-b, respectively.

Generating $C$ requires a delayed version, $ck_\Delta$, of the clock signal. The trivial implementation of a delay element is a chain of an even number of suitably sized inverters. Unfortunately, this implementation is highly power and area inefficient when the delay is much larger than a single inverter delay. Several implementations have been proposed that overcome this limitation. We adopted the delay element proposed in [8], whose power dissipation is approximately three times that of a single inverter and approximately four times larger (when the delay is approximately 16 times that of a single inverter).

The $C$ signals are distributed through a dedicated network (made of tapered buffers and delay elements) that stems from the clock tree in a single point. Since the tap point is a minimum-size buffer, the additional load on the clock tree is usually negligible.

### 3 Gate Selection and Clustering

The basic version of the gate freezing algorithm is shown in Figure 4. The procedure receives as inputs a mapped, placed and routed circuit, $M$, a library of cells, $L$, containing both ordinary cells and the corresponding $F$-Gates, and the maximum number $NFG$ of gates to be selected for replacement with $F$-Gates. The procedure first performs static timing analysis (Line 1) and power estimation (Line 2); then, it selects candidate cells for replacement with $F$-Gates (Line 3). Finally, it replaces each selected gate $g$ with the corresponding $F$-Gate (Line 4), it computes the timing for the control signal of each gate (Line 5), and it instantiates the circuitry for generating them (Line 6). We first analyse procedure SelectCandidates (Section 3.1), that returns the set of gates to be replaced by $F$-Gates.

Then, we focus our attention on gate clustering (Sections 3.2 and 3.3), a key step that is required to make gate freezing applicable in practice.

#### 3.1 Selection of the Target Cells

The selection of the cells to be replaced with $F$-Gates is mainly driven by the amount of glitching observed at the output of each cell. Additionally, we take into account the capacitive load of the nodes. Therefore, our simple procedure for cell selection is to sort the network nodes in decreasing order of glitching activity weighted by load capacitance (we call this metric glitching-capacitance product, $gc$ for brevity), and to select the first $NFG$ gates of the sorted list. Glitching is measured by counting the number of transitions at the output of each gate obtained by accurate (i.e., real delay) simulation, and by subtracting from this amount the number of transitions obtained through ideal (i.e., zero-delay) simulation. The latter values must either be 1 or 0, depending on whether there is a transition or not. This difference represents the spurious transitions propagating through a gate.

Remember that nodes with zero slack cannot be selected, because replacing them with $F$-Gates would slow down the circuit; consequently, node selection is applied only to gates with non-zero slack. In addition, we observed that picking new gates in the recursive function of already chosen gates may be disadvantageous. In fact, the presence of an $F$-Gate in the fanin cone of a gate tends to reduce the glitching activity of the gate itself. This effect could be taken into account by performing power analysis after selecting each node; however, this process could make the optimization procedure very slow in the case of large circuits.
We have adopted an enhanced selection procedure that
approximately takes into account previous selections. Ini-
tially, all nodes are marked with an integer label, which
is set equal to 0. Then, the node with highest glitching-
capacitance product is selected. If two or more nodes have
the same gc value (with a 10% tolerance margin), we use
the total capacitive load of the transitive fanout as a tie-
breaker. The gate with highest total capacitive load of the
transitive fanout is chosen first. The label of all the nodes
in the transitive fanout of the selected node is then incre-
mented. Node labels are used as tie-breaker for successiv-
e selections. If two or more nodes have the same gc value
(with the usual 10% tolerance margin), we select the one
with smallest label. If this rule is not sufficient to break all
the ties, the total capacitive load of the transitive fanout is
used. If the ties persist, we choose randomly. The pro-
cess is stopped after NFG nodes have been selected.

3.2 Clustering

In the algorithm of Figure 4, one control signal for each
selected gate needs to be generated. Although the num-
ber of gates replaced is usually a small fraction of the cells
that are present in the circuit implementation (less than
5%), the generation of a control signal for each F-Gate
is still quite impractical, since each signal has generally a
duty cycle value different from any other, and there is then
little chance to exploit some sharing.

To reduce both the area of the control circuitry imple-
mentation and the routing of the control signals, we need to
limit the number of such signals as much as possible. One
way of doing this is to cluster the F-Gates according to the
values of their arrival times Δ's, as defined in Equation 1.

The clustering problem can be stated as follows: Given a
set of NFG gates G = \{g1, ..., gNFG\}, and their arrival
times Δg1, ..., ΔgNFG, find a partition \(\mathcal{P} = \{P1, ..., PK\}\)
of the Δ's such that: i) The partition is balanced; ii) The
variance of the Δ times within each block \(Pi\) of the
partition is minimized; iii) The number of blocks, \(K\), is
small.

The partition \(\mathcal{P}\) on the arrival times induces a partition
on the set of selected gates \(\mathcal{G} = \{G1, ..., GK\}\).

The requirement of the minimum variance is related to the
error that this approximation introduces. In fact, all
the gates belonging to the same block \(Gi\) will be fed by
the same control signal, whose delay, with respect to the
clock, is determined by the earliest of their Δ times. More
formally, the delay \(DAi\), for all the control signals of the
gates in \(Gi\) is given by:

\[
DAi = \min_{g \in Gi} \Delta(g) \tag{2}
\]

where \(\Delta(g)\) is defined as in Equation 1.

The approximation introduced by clustering some of the
control signals together arises from the fact that all the
gates g in \(Gi\) having \(\Delta(g)\) larger than the \(DAi\) will allow
the propagation of the glitches occurring in the time in-
terval \([\Delta(g_{min}), \Delta(g)]\), where \(g_{min}\) is the gate in \(Gi\) that
determines the bound of Equation 2.

Figure 5 shows the pseudo-code of the clustering-based
gate freezing algorithm.

![Figure 5: Clustered Gate Freezing Algorithm.](image)

The I/O interface and the flow of this algorithm are similar
to those of Figure 4. The main difference stands in Lines 4-
8, where a proper partition of the candidate gates is first
built (Line 4) (the details on the clustering procedure are
described in Section 3.3). Procedure Clustering returns a
partition \(\mathcal{P} = \{P1, ..., PK\}\) of the Δ times, and the cor-
responding partition of the selected gates \(\mathcal{G} = \{G1, ..., GK\}\).

Then, the algorithm iterates over the \(K\) blocks of \(\mathcal{G}\) and
derives, for each cluster \(Gi\), the values of \(DAi\), as defined
by Equation 2 (Line 5). Each gate \(g \in Gi\) is then re-
placed with the corresponding \(F\)-Gate (Line 6). After all
the gates in \(Gi\) have been replaced, the circuitry for gen-
erating the shared control signal is determined (Line 7) and
incrementally added to the circuit layout (Line 8).

3.3 Clustering Heuristics

The clustering problem, as stated in Section 3.2, is too
general. In fact, the solution with optimal cost may require a
number of clusters, \(K\), which may be too large. We have
implemented two heuristics that solve the clustering prob-
lem when \(K\) is upper-bounded by a user-specified value.

Experiments have shown that the first one, based on clus-
ters growth, works better on clusters of small cardinali-
ty (and thus on small circuits), while the second heuristics,
based on matching, best performs on mid to large-sized
clusters.

In both cases, the cost function \(e_{avg}\), we have used to drive
the clustering procedure is the average cluster variance
[i.e., the average Δ of the gates in \(Gi\)], whose formal def-
nition is the following (we assume to have a total of \(K\)
clusters, \(\{G1, ..., Gi, ..., GK\}\)), whose formal defi-
nition is the following (we assume to have a total of \(K\)
clusters, \(\{G1, ..., Gi, ..., GK\}\)),

\[
e_{avg} = \frac{\sum_i \sigma_i}{K}
\]

where:

\[
\sigma_i = \sum_{g \in Gi} (\Delta(g) - \Delta_{avg})^2
\]

is the variance of the Δ values in cluster \(Gi\). Obviously,
better solutions are identified by lower values of \(e_{avg}\).

The details of the two heuristics mentioned above are omit-
ted for space reasons; for further information, the inter-
ested reader can refer to [9].
Table 1: Experimental Results.

<table>
<thead>
<tr>
<th>Circ</th>
<th>PI</th>
<th>PO</th>
<th>Gates</th>
<th>Original</th>
<th>Optimized</th>
<th>Optimization</th>
<th>Glitch</th>
<th>GPC</th>
<th>PC</th>
<th>Area</th>
<th>Glitch Power Reduction</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15B</td>
<td>1</td>
<td>0</td>
<td>200</td>
<td>2.42</td>
<td>2.42</td>
<td></td>
<td>0.00</td>
<td>0.0</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>25B</td>
<td>1</td>
<td>0</td>
<td>200</td>
<td>2.42</td>
<td>2.42</td>
<td></td>
<td>0.00</td>
<td>0.0</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>15A</td>
<td>1</td>
<td>0</td>
<td>200</td>
<td>2.42</td>
<td>2.42</td>
<td></td>
<td>0.00</td>
<td>0.0</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>25A</td>
<td>1</td>
<td>0</td>
<td>200</td>
<td>2.42</td>
<td>2.42</td>
<td></td>
<td>0.00</td>
<td>0.0</td>
<td>0</td>
<td>0.00</td>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>

4 Experimental Results

We have implemented the clustered gate freezing procedure using SIS as the gate-level front-end. A pictorial description of our experimental environment is shown in Figure 6.

Figure 6: Experimental Environment.

The initial circuits were optimized using scripts delay, and mapped using map = 1 -2F onto a gate library consisting of two and three input NAND and NOR gates, and INV and BUF gates with three different driving capabilities. Placement and routing of the tech-mapped circuits onto a 0.6μm static CMOS physical-level library was done using Alliance [10]. Gate-and switch-level netlists were extracted from the layout using an in-house tool; such netlists were used for gate-level timing analysis (using SIS) and for switch-level power simulation (using Insim [11]). After applying the clustered gate freezing procedure, the layouts were incrementally modified using Alliance, and the power dissipated by the optimized circuits was estimated (using Insim) on the switch-level netlists extracted from the final layouts. Timing verification was also performed using SIS on the gate-level netlist derived from the final layout.

The circuits we considered for the experiments are the Iocc's5 benchmarks [2]; results for some of the largest Mcnc'91 multi-level networks are also reported [3].

Table 1 summarizes the data. Columns GP, TP, and Area give the glitch power (in mW), the total power (in mW), and the layout area (in Λ²) before (columns Original) and after (columns Optimized) optimization. Columns [%] give the percentages of power and area variation. The three right-most columns indicate the clustering heuristics, the number of clusters [NC], and the number of frozen gates [NFG] that have been chosen for each experiment.

Results, though preliminary, are very promising. In fact, an average glitch power reduction of 14.0%, resulting in an average reduction of the total power of 6.3% has been achieved at the cost of a negligible area increase (2.8% on average). Obviously, the speed of the circuits has not changed, since only non-critical gates have been replaced with the (marginally slower) F-Gates.

Note that gate freezing exclusively targets glitch power reduction. Therefore, the quality of the results would only be marginally affected by gate-level power optimization techniques (e.g., POSE [12]) that minimize zero-delay power.

5 Conclusions

We have proposed a technique for glitch power minimization in combinational circuits. The method consists of transforming some high-glitching gates into modified devices that are able to filter out spurious output transitions whenever a proper control signal is activated. The technique has the distinctive feature of being applicable as a post-layout processing step, since it performs shunt optimization on the placed and routed description, and it only requires incremental re-wiring.

References