On optimal flip-flop grouping for VLSI power minimization

Shmuel Wimer*
Bar-Ilan University, Engineering Faculty, Ramat-Gan, Israel
Technion-Israel Institute of Technology, EE Department, Haifa, Israel

A R T I C L E   I N F O
Article history:
Received 13 December 2012
Received in revised form 7 June 2013
Accepted 7 June 2013
Available online 14 June 2013

Keywords:
NP-hardness
Minimum cost perfect matching
VLSI clock gating
VLSI power optimization

A B S T R A C T
Data-driven clock gating is reducing the total power consumption of VLSI chips by 20%. There, flip-flops are grouped and share a common clock signal. Finding the optimal clusters is the key for maximizing the power savings. Clustering by the minimal cost perfect graph matching algorithm (MCPM) proposed by other works is not optimal. We show that the optimal clustering problem is NP-hard, and study the quality of MCPM heuristics, showing by experiments that it falls 5% above the optimal solution.

1. Introduction

One of the major switching power consumers in computing and consumer electronics products is the system’s clock signal, typically responsible for 30% to 70% of the total switching power consumption [11]. Several techniques to reduce the switching power have been developed, of which clock gating is predominant. Ordinarily, when a logic unit is clocked, its underlying sequential elements (flip-flops) receive the clock signal regardless of whether or not they will toggle in the next cycle. With clock gating, the clock signals are conditioned (ANDed) with explicitly predefined enabling signals. Clock gating is employed at all chip levels: system architecture, block design, logic design and gates [2,8]. Several methods to take advantage of this technique are described in [3,6,13], with all of them relying on various heuristics in an attempt to increase clock gating opportunities.

This paper studies data-driven clock gating, employed for flip-flops (FFs) at the gate-level, which is the most aggressive possible. The clock signal driving an FF is disabled (gated) when the FF’s state is not subject to change in the next clock cycle [5]. In a recent study, a model for data-driven gating was developed based on the toggling (switching) activity (state changes) of the constituent FFs [15]. The optimal fan-out of a clock gater yielding maximal power savings was derived based on the average toggling statistics of the individual FFs. While [15] answered the question of what is the group size that maximizes power savings, this work studies the problem of which FFs should be placed in a group to maximize the power reduction and how to algorithmically derive those groups. We subsequently use the terms activity, toggling and switching interchangeably.

A data-driven clock gating circuit is illustrated in Fig. 1. By XOR-ing its output with the present input data that will appear at its output in the next clock cycle, an FF checks whether its state is subject to change, thus finding out whether its clock can be disabled in the next cycle. The outputs of k XOR gates are ORed and then latched to generate a joint gating signal for k FFs. The combination of a latch with AND gate is called Integrated Clock Gate (ICG) [10], commonly used by commercial electronic design automation (EDA) tools. Notice that a single ICG is amortized over k FFs. There is a clear trade-off between the number of saved (disabled) clock pulses and the hardware overhead. With an increase in k the hardware overhead decreases, but so does the probability of disabling, obtained by OR-ing the k enable signals.

The FFs of a system need to be clustered in k-size sets such that the power savings will be maximized. Grouping FFs for joint clock gating was described in [12] as a part of physical layout synthesis. While addressing design factors as clock-skew, power, and area minimization, it was not aware of the toggling correlations of the underlying FFs, which this paper does. The optimal value of k was obtained by [15] under worst-case assumption of FFs toggling independence. In reality however, the toggling is correlated, so one can expect for higher savings than the theoretical lower bound obtained under independence assumption. In the sequel we use the terms grouping and clustering interchangeably.

The next section presents the problem of optimal FF grouping by introducing a graph model followed by a formulation of the
associated optimization problem. Section 3 shows the inherent difficulty of the problem and proves its NP-hardness. Section 4 describes a heuristic grouping algorithm.

2. Optimal FFs grouping for joint clock gating

Let n FFs be clocked during m + 1 cycles. A first step towards an optimal FFs grouping is to take advantage of the correlations of their toggling. Let \( a = (a_1, \ldots, a_n) \) be the activity of an FF, where \( a_t = 0 \), \( 1 \leq t \leq m \), if the FF stays unchanged (no toggling) from time \( t-1 \) to time \( t \), and \( a_t = 1 \), otherwise. The term \( \|a\| = \sum_{t=1}^{m} a_t \) is proportional to the power consumed by the FF’s switching. All the \( n(n-1)/2 \) pairs \( (a_i, a_j) \), \( 1 \leq i < j \leq n \), are bit-wise XORed to yield the number \( \|a_i \oplus a_j\| \) of redundant clock pulses occurring if \( FF_i \) and \( FF_j \) are jointly clocked by a common gater. A key consideration in selecting FFs to be driven by a common gater is their activity similarity given by \( \|a_i \oplus a_j\| \). The smaller it is, the more desirable it is to jointly clock FF \( i \), and FF \( j \).

To model the switching power consumed when driving FFs pairs \( (k = 2) \) with a common clock gater, an \( n \)-vertex complete weighted graph \( G(V, E, w) \), called FF pairwise activity graph, is defined. Assume w.l.o.g that \( n \) is even (we could otherwise add a new toggling artificial FF and set to zero the weight of its entire incident edges.) A complete \( k \)-uniform hypergraph \( H(V, E, w) \), called FF grouping activity hypergraph, is defined, where for a subset \( V \subseteq V \) and \( |V| = k \), \( \epsilon_V = \{v_i\}_{i \in V} \in E \) defines a hyper edge. It follows that \( |E| = \binom{n}{k} \). A hyper edge \( \epsilon_V \) is associated with a joint activity vector \( \sum_{a \subseteq E} a \), defined by the bit-wise ORing of the \( k \) toggling vectors. A hyper edge \( \epsilon_V \) is assigned a weight \( w(\epsilon_V) = \sum_{V \subseteq V} \|a_i \oplus a_j\| \).

The first sum on the right hand side of (1) is an essential power component charged to the toggling of the individual FFs, and is independent of the pairing. Therefore, to consume minimum switching power (or alternatively, achieve maximum switching power savings) it is necessary to minimize \( \sum_{\epsilon \in E} w(\epsilon) \), which turns into the well-known minimal cost perfect graph matching (MCPM) problem, for which polynomial complexity algorithms are known [9].

The extension for \( k > 2 \) is straightforward. Assume w.l.o.g that \( n \) is divisible by \( k \). (We could otherwise add a new toggling artificial FFs and set to zero the weight of their entire incident edges.) A complete \( k \)-uniform hypergraph \( H(V, E, w) \), called FF grouping activity hypergraph, is defined, where for a subset \( V \subseteq V \) and \( |V| = k \), \( \epsilon_V = \{v_i\}_{i \in V} \in E \) defines a hyper edge. It follows that \( |E| = \binom{n}{k} \). A hyper edge \( \epsilon_V \) is associated with a joint activity vector \( \sum_{a \subseteq E} a \), defined by the bit-wise ORing of the \( k \) toggling vectors. A hyper edge \( \epsilon_V \) is assigned a weight \( w(\epsilon_V) = \sum_{V \subseteq V} \|a_i \oplus a_j\| \).

which is the total number of redundant clock pulses incurred by clocking the \( k \) FFs corresponding to \( \epsilon_V \) with a common gater.

Let \( E' \subseteq E \) be an exact cover of the vertices of \( H(V, E, w) \) by \( n/k \) hyper edges (a vertex belongs to one and only one hyper edge). The total power \( P \) consumed by the clock signal depends on the total number of pulses driving the FFs, given by

\[
P = \sum_{\epsilon \in E'} k \|a_i \oplus a_j\| = \sum_{\epsilon \in E'} \|a_i \oplus a_j\| + \sum_{\epsilon \in E'} w(\epsilon_V).
\]
However, the optimal 4-size group is (FF1, FF2, FF6, FF7) and (FF3, FF4, FF5, FF8), yielding 35 redundant clock pulses. The pairs (FF2, FF6) and (FF7, FF8) have been split between the two 4-size sets shown in Fig. 2(b). Consequently, the optimal solution could not be obtained by a repetitive MCPM. This follows from the inherent difficulty of the MIN_CLK_GATE problem, shown next to be NP-hard.

We prove the NP-hardness of MIN_CLK_GATE by a polynomial reduction of the X3C exact covering problem [7] into a decision problem version of MIN_CLK_GATE. Minimizing \( \sum_{v \in S'} \nu(v) \) in (3) can equivalently be viewed as maximizing the number of clock pulses saved by group FFs, where an FF that is not subject to toggling at the next clock cycle does not receive a clock pulse. Consider the following version of the decision problem:

**Problem:** MIN_CLK_GATE

**Instance:** A set of n FFs where n is divisible by k. Each FF is associated with a toggling vector \( a = (a_1, a_2, \ldots, a_m) \) during \( m+1 \) clock cycles, where \( a_i = 1 \) if the FF changes states between clock cycles \( j-1 \) and \( j \), \( 1 \leq j \leq m \), and \( a_i = 0 \) otherwise.

**Question.** Is there a grouping of the n FFs in \( n/k \) k-size disjoint sets such that savings of at least \( n/k \) clock pulses are achieved by jointly driving the FFs in each set with a common clock gate?

**Theorem.** The MIN_CLK_GATE decision problem is NP-complete.

**Proof.** We will consider a special case of MIN_CLK_GATE where n is divisible by k = 3 and in every clock cycle exactly \( n/3 \) FFs are toggling. The implied decision problem is whether there exists a grouping of the n FFs in 3-size sets, resulting in savings of \( n/3 \) clock pulses at least. The toggling vectors of the FFs during the \( m+1 \) clock cycles are represented by a \( n \times m \) matrix \( T = (t_{ij}) \) defined as follows:

\[
t_{ij} = \begin{cases} 
1 & \text{FF is toggling from } j-1 \text{ to } j \text{ clock cycle} \\
0 & \text{otherwise,} 
\end{cases}
\]

\( 1 \leq i \leq n, \ 1 \leq j \leq m. \) (4)

Let the 3-size set \{FF1, FF2, FF3\} be driven by a common clock gate, and let \( T_{(1,2,3)} \) be its corresponding \( 3 \times m \) submatrix of \( T \). There can be any number, from zero to three, of 1s in a column of \( T_{(1,2,3)} \). If all elements are 0, ORing the 3 os yields a 0 enabling signal and the gate will not send a clock pulse to any of \{FF1, FF2, FF3\}, hence three pulses are saved. In all other case at least one of the three FFs toggles and must therefore be clocked. An ORing results in a joint enabling signal of 1 and all the three FFs will receive a clock pulse, thus no savings occur. Once the FFs are grouped in 3-size sets, calculating the implied savings is straightforward and takes \( O(mn) \) time, showing that MIN_CLK_GATE decision problem is NP.

We will show next a polynomial reduction of the X3C decision problem [7] into the MIN_CLK_GATE decision problem. Recall that a X3C asks whether for a given set \( S \) of subsets, \( S = \{S_1, \ldots, S_m\}, |S| = 3, 1 \leq j \leq m \), of a universal set \( U = \{U_1, \ldots, U_n\} \), there is a subset \( S' \subseteq S \) such that \( \bigcup_{S \in S'} S = U \) and \( \bigcap_{S \in S'} S = \emptyset \) for any \( S_i \in S' \) and \( S_j \in S, i \neq j \). We may restrict ourselves to instances of the X3C cover problem where the subsets in \( S \) are unique.

A transformation \( f \) of a X3C instance into MIN_CLK_GATE is defined as follows. An element \( U_i \in U, 1 \leq i \leq n \), is associated with an FF, and a subset \( S_i \in S, 1 \leq j \leq m \) is associated with a clock cycle. An \( n \times m \) zero–one toggling matrix \( T = (t_{ij}) \) is thus defined by:

\[
t_{ij} = \begin{cases} 
0 & \text{if } U_i \in S_j \\
1 & \text{otherwise,} 
\end{cases}
\]

\( 1 \leq i \leq n, \ 1 \leq j \leq m. \) (5)

Since \( |S| = 3 \), a column in \( T \) has exactly three 1s. Let the answer to MIN_CLK_GATE be yes. There is a grouping \( S', |S'| = n/3 \), of the n FFs into disjoint sets of 3 FFs each, yielding savings of \( n/k \) clock pulses at least. Consider a set \{FF1, FF2, FF3\} \( \in S' \) and its corresponding \( 3 \times m \) sub matrix \( T_{[g,r]} \subseteq T \), representing their joint toggling during the \( m+1 \) clock cycles. Let \( p (FF1, FF2, FF3) \) denote the number of clock pulses saved in jointly clocking \{FF1, FF2, FF3\} during the \( m+1 \) clock cycles. The yes answer to MIN_CLK_GATE means that:

\[
\sum_{S'} p (FF1, FF2, FF3) \geq n. \]

\( (6) \)

We subsequently show that in (6) only the equality can hold. Let \( T', 1 \leq j \leq m \), denote a column of \( T \), and let \( p (T') \) be the number of clock pulses saved in a transition from clock cycle \( j-1 \) to \( j \). By the construction in (5) \( T \) has exactly three 1s in every column, and since in \( S = \{S_1, \ldots, S_m\} \) of the X3C problem all \( S_i \) are distinct, at most one column of \( T_{[g,r]} \) has three 1s. The MIN_CLK_GATE grouping \( S' \) comprises \( n/3 \) 3-size FFs sets, so there are at most \( n/3 \) columns of \( T \) where all three 1s fall within a set in \( S' \). Consequently:

\[
\sum_{j=1}^{m} p (T_j) \leq 3 \times n/3 = n. \]

\( (7) \)

It follows from inequalities (6) and (7) that a yes answer to MIN_CLK_GATE implies total savings of exactly \( n \) clock pulses. These savings are obtained from the \( n/3 \) FFs sets, which from the transformation in (5) uniquely defines \( n/3 \) disjoint sets \( \{U_{i1}, U_{i2}, U_{i3}\} \), hence yielding a yes answer to the X3C problem. The converse direction where a yes answer to X3C implies a yes answer to MIN_CLK_GATE is straightforward. We associate every FF, with an element \( U_i \in U \), and its toggling in a transition from clock cycle \( j-1 \) to \( j, 1 \leq j \leq m \), with the definition of (5); A yes answer to X3C implies \( n/3 \) FFs sets satisfying (6), hence yielding a yes answer to the MIN_CLK_GATE decision problem. \( \square \)

**4. Results of iterative MCPM heuristic**

We have shown that MIN_CLK_GATE is NP-hard. The MCPM heuristic to solve MIN_CLK_GATE is still practical, yielding results close to the minimal cost that can be obtained by SPP solution, as demonstrated by the experimental result. Since the number 3\(^m\) of SPP variables explodes with the number n of FFs and the group size k, we could afford a comparison of only a small case
of \( n = 94 \) FFs, taken from a real VLSI design. The FF toggling benchmark spans \( m = 10^5 \) clock cycles and has average toggling \( p = 0.0736 \). The group size \( k = 4 \) is appropriate for the given toggling probability [15], resulting in a minimum cost SPP with \( \binom{94}{4} \approx 3.05 \times 10^6 \) variables and 94 constraints. Its solution was compared to the results obtained by MCPM heuristic. (Recall that for \( k = 2 \) both MCPM and SPP yield by their definition the same minimum.) The absolute minimum obtained by the minimum cost SPP algorithm has \( \sum_{e \in E} w(e) = 578,671 \) redundant pulses, while the MCPM algorithm yields 604,545 redundant pulses, which is 4.47% above the optimal solution.

The MCPM algorithm has also reasonable run time performance as shown in Table 1, obtained from a design comprising \( n = 4.9 \times 10^5 \) FFs. The toggling benchmark spans \( m = 10^5 \) clock cycles and has \( p = 0.05 \) average FF toggling. The experiment ran on a 2 GHz processor with 2 GB RAM. Since all FFs pairs are allowed, the pairwise activity graph includes \( 4.9 \times 10^5 \) vertices and \( n(n - 1)/2 \approx 1.2 \times 10^7 \) edges. Due to FFs placement and proximity constraints resulting from VLSI design considerations (whose discussion is beyond the scope of this paper), the size of such a graph in practice is much smaller. Groups of \( k = 2^k \) were tested, \( k = 2, 4, 8, \ldots, 128 \) have been examined. The results are summarized in Table 1. There, the number of redundant clock pulses as obtained by (3) is far smaller than that implied by a random grouping, yielding \( mn \left[ 1 - p - (1 - p)^k \right] \) redundant pulses. The low number of redundant pulses obtained by the MCPM heuristic compared to random grouping stems from the correlations of FFs’ activities which the grouping algorithm took advantage of. The run-time growth is nearly logarithmic in \( K \). This follows from the iterative nature of group constructions where at each step a problem half the size of the former iteration is solved.

Acknowledgments

The author wishes to thank the Technion VLSI Lab for developing the FF grouping software, the Israel Ministry of Industry (MAGNET program) for supporting this work, and Dr. Osnat Keren for providing the counter example. He is also grateful for the corrections and useful comments made by the reviewer.

References