FAST TOGGLE RATE COMPUTATION FOR FPGA CIRCUITS

Tomasz S. Czajkowski  
Department of ECE, University of Toronto,  
Toronto, Ontario, Canada  
czajkow@eecg.utoronto.ca

Stephen D. Brown  
Department of ECE, University of Toronto,  
Toronto, Ontario, Canada  
brown@eecg.utoronto.ca

ABSTRACT
This paper presents a fast and scalable method of computing signal toggle rate in FPGA-based circuits. Our technique is a vectorless estimation technique, which can be used in a CAD tool to identify the parts of the circuit that can benefit from power optimization. A key advantage of our approach is its ability to efficiently account for spatial correlation of related logic cones, which is accomplished using a novel XOR-based decomposition. In addition, our approach uses post-routing circuit delays to account for glitches in a logic circuit.

The proposed approach was tested on 14 MCNC benchmark circuits compiled for the Altera Stratix II devices. The results indicate that our method improves the vectorless estimation technique available in the latest version of Altera’s Quartus II commercial CAD tool, reducing the average error by 37% and standard deviation by 59%.

1. INTRODUCTION
Field-Programmable Gate Array (FPGA) devices are a popular choice for low to medium volume digital applications. They can implement various different circuits without a need to fabricate a new chip, reducing costs and time to market. This flexibility is paid for by increased power dissipation. The power consumed by a circuit implemented on an FPGA device in a 90nm process, or smaller, presents a problem, especially if an FPGA device is to be used in a mobile application.

The power dissipation of a 4-input LUT based FPGA consists 40% of static power and 60% of dynamic power, with static power increasing with shrinking feature sizes and increasing LUT sizes [3]. The dynamic power dissipation is determined by the toggle rate of signals in the circuit, which can as much as double when glitches are present [2].

Several works that address the topic of toggle rate estimation already exist [1][2][4][6-13]. In the work on transition density [1] the toggle rate is represented using the theory of probability. The main advantage of this approach is that toggle rate can be represented using statistical analysis, producing reasonably accurate results for the overall power dissipation in a logic circuit. However, the accuracy of results on a per connection basis is low, preventing power reduction algorithms from effectively optimizing a logic circuit.

The work in [2] looks at the transition density (average number of times a signal toggles per unit time) [1] and how this concept applies to circuits implemented on FPGAs. The transition density is represented as a weighted sum of transitions generated and propagated through a LUT. The generated transitions are a function of LUT depth and the number of paths to LUT inputs. Propagated transitions are due to glitches that propagate through a LUT and are based on the logic function the LUT implements.

The work in [2] was intended to predict the toggle rate of a circuit before it is placed and routed. The work in [4] focuses on transition density computation once circuit delays are known. Their approach is in essence an event driven simulation, except that logic transitions are represented using a probabilistic waveform. This method effectively combines multiple events on primary inputs of a circuit into a single probabilistic waveform to obtain transition density values.

A probabilistic model introduced in [7] accounts for both temporal and spatial correlation in a circuit. Using ROBDDs to perform correlation analysis improved the computation time of their method, and the approach scales much better in terms of processing time than competing approaches presented in [8] and [13]. Unfortunately, the method only computes the static and transition probabilities, ignoring glitches caused by delay imbalance in a circuit. In FPGA circuits such glitches can cause the number of transitions on any given wire to double [2], resulting in an increased power dissipation. Hence, a toggle rate analysis method needs to account for glitching as well as correlation of logic signals.

This work presents a fast approach to compute toggle rate on a per wire basis using an XOR-based logic decomposition. It considers temporal and spatial correlations, as well as glitching of logic signals due to delay imbalance in a logic circuit. The proposed approach processes the circuit in topological order computing transition density for each LUT. For LUT inputs that are correlated, the approach performs a Gaussian Elimination based decomposition. It first determines the cones of logic that capture the dependency between these inputs and then decomposes these cones into subfunctions that expose the correlation between them in an efficient manner. The decomposition is then used to account for spatial correlation of LUT inputs.

The approach was tested on several MCNC benchmarks as well as several industrial benchmarks, targeting the Altera Stratix II FPGA. The results show a significant improvement of the approach currently available in the most recent release of Quartus II.

The remainder of this paper is organized as follows:
Section 2 contains background information. Section 3 describes our toggle rate analysis method, focusing on temporal and spatial correlation. The equations used to compute correlation are derived in Section 4, while the experimental results are presented in Section 5. The conclusion and avenues for further research are in Section 6.

2. BACKGROUND

The average dynamic power dissipation of a logic circuit is defined by the following equation:

\[
P_{\text{avg}} = \sum_{\text{all sets } i} \left( f_i C V^2 \right)
\]

where \( f_i \) is the transition density measured on a per second basis, \( C \) is the net capacitance, and \( V \) is the power supply voltage. In a circuit where the primary inputs are registered, we can expand \( f_i \) as \( s_i f_{\text{clock}} \), where \( f_{\text{clock}} \) is the clock frequency and \( s_i \) is the per cycle transition density of a net. The transition density is a key parameter in this equation as it changes depending on the logic function of a component that drives it, as well as logic functions in its transitive fanin, logic and wire delays, as well as glitches.

The computation of transition density involves the use of several parameters: static probability, 0–1 transition probability and 1–0 transition probability. Static probability defines the likelihood that a logic signal is at a logic value of 1. The 0–1 transition probability determines how likely a signal is to change state once it is at a logic 0 value. Similarly, the 1–0 transition probability accounts for the likelihood of a signal changing state when it is at a logic 1 value. The latter two parameters are commonly referred to as the temporal correlation of a signal. While these three parameters are necessary to compute transition density for any given logic function, each of them is affected by what is known as spatial correlation [5].

Spatial correlation is a measure of the relationship between logic signals at an input of a logic gate. This occurs when logic expressions for each of the inputs are related by common set of signals, either primary inputs or an internal signal. In both cases a particular logic state of one input affects the static probability of the other inputs.

To compute spatial correlation for any given set of gate inputs, it is important to be able to quickly determine the nature of the relationship between these signals. To do this, we utilize an XOR-based logic decomposition based on Gaussian Elimination. The basic idea behind this decomposition is demonstrated on the example in Figure 1.

Figure 1 shows a simple logic function of four variables. We can see that each column of the truth table is different from the others, however a close inspection reveals that column 11 is an XOR of columns 01 and 10. By utilizing this fact we can extract two logic sub functions, \( G_1 = c \) to represent column 01 and \( G_2 = d \) to represent column 10. Column 11 would therefore be \( G_1 \oplus G_2 \), allowing the overall function to be synthesized as \( f = b G_1 \oplus a G_2 = bc \oplus ad \). The problem of finding such relationships between logic columns is solved by using Gaussian Elimination. By utilizing Gaussian Elimination along with BDDs the approach can be effectively applied to large circuits and the synthesis of multi-output functions [14].

3. TOGGLE RATE COMPUTATION

The computation of the toggle rate of a logic signal in an FPGA circuit consists of two components - transitions generated by a LUT and glitches propagated from LUT inputs to its output.

The transitions generated by a LUT are those that occur due to changes in logic state of LUT inputs. If LUT inputs change at different times, then glitches may be generated. For example, a 3-LUT with inputs \( a, b, \) and \( c \) can produce two intermediate values while its inputs change from input pattern 000 to 111. Thus if the LUT produces a logic 0 for inputs 000, logic 1 for 001, logic 0 for 011 and logic 1 for inputs 111, then the LUT output will show 3 transitions, whereas only one should occur. Furthermore, these glitches may be propagated by LUTs downstream. Transitions due to glitches would occur on top of the transitions generated by a LUT and should be added to the toggle rate for a given LUT output.

In the following explanations we utilize the concept of transition density to express the toggle rate, measuring transitions on a per clock cycle basis.

3.1 Basic Approach

The basic approach is to process each LUT one at a time in a topological order. For each LUT, we evaluate the static probability, transition probability, transition density, 0–1 and 1–0 transition probabilities and use them in computations downstream. This is done by computing transitions generated by, and propagated through, a LUT.

Transitions generated by a LUT are computed using the static probability for each possible input valuation, the probability of transition from one valuation to another, and the arrival time of each signal at a LUT input. The transition density generated by a LUT is computed as follows:

\[
D_s = \sum_{i=0}^{1} \sum_{j=0}^{1} \left( P[i] P[j] Trans(i, j) \right)
\]

where \( i \) and \( j \) represent input valuations, and \( P[i] \) is the static probability of a given input valuation, \( P[j] \) is the probability...
that inputs change from pattern \( i \) to \( j \) and \( \text{Trans}(i,j) \) is the number of transitions generated when input pattern changes from \( i \) to \( j \). The last term uses the arrival time of each input signal to determine the number of transitions that will occur at the output of a LUT by changing input pattern \( i \) to \( j \) one bit at a time in the order determined by their arrival time.

Now that the generated transitions are accounted for, we need to compute the transition density component due to glitches propagated through a LUT. To account for the glitches we utilize the concept of Boolean Difference in a fashion similar to [1]. However, we impose a constraint on glitch forwarding such that the glitches from an input \( x \) of a LUT are forwarded to its output only during such input transitions that both the initial and the final input valuations are sensitive to changes on input \( x \). Because of this restriction, we need to account for the probability of a particular input transition in the same manner as in Equation 2. Thus, the probability glitches on input \( x \) are transferred to the LUT output is computed as follows:

\[
P_{\text{glitch}}(i, j) = \frac{df(i)}{dx} \cdot \frac{df(j)}{dx} \cdot P[i] \cdot P[j|i]
\]

where \( f \) is the function implemented by the LUT and \( df/dx \) is the Boolean Difference of \( f \) with respect to LUT input \( x \). The above equation will be equal to \( P[i]P[j|i] \) only when both the initial (i) and final (j) input valuations cause function \( f \) to change value if input \( x \) changes value (\( df/dx=1 \)). Otherwise, the equation will evaluate to 0, signifying that glitches from input \( x \) will not propagate to the LUT output during the specified transition. Using this definition we can express the transition density due to glitching on a LUT input \( x \) as:

\[
D_p(x) = \sum_{i=0}^{1} \sum_{j=0}^{1} \left( P_{\text{glitch}}(i, j) \cdot D(x) - P_{i}(x) \right)
\]

where \( P_i(x) \) is the transition probability of input \( x \) and \( D(x) \)-\( P_i(x) \) is the average number of glitches per cycle on input \( x \).

Using the above approach we can define the total transition density for any given wire \( y \) as:

\[
D(y) = D_g + \sum_{i=1}^{n} D_p(x_i)
\]

While the above computation is relatively simple, its key parameters (\( P[i] \) and \( P[j|i] \)) are affected by correlation [5].

### 3.2 Temporal Correlation

The temporal correlation describes the behaviour of a logic signal with respect to its previous value. It is defined by the 0→1 and 1→0 transition probabilities. These probabilities determine how likely a signal is to change state given its present state. For example, a signal that assumes a logic value 1 for a single clock cycle once every 8 cycles on average, would have a 0→1 transition probability of 1/8, while its 1→0 transition probability would be equal to 1. In our model these probabilities affect the \( P[j|i] \) parameter, which determines the likelihood of a transition from one state to another.

### 3.3 Spatial Correlation

The spatial correlation is a more complex type of correlation observed in logic circuits. It is important because it can change both the probability of any given input pattern as well as the probability of input transition.

The basic idea behind spatial correlation is that two, or more, signals may share logic in their transitive fanin. As such, their logic state can be related to a varying degree. An extreme example is one where two inputs, \( x \) and \( y \), are related by inversion, in which case either \( xy=00 \) or \( xy=11 \) are possible input valuations.

Similarly, the spatial correlation may change the probability of input transition. For example, suppose that two inputs to a LUT, \( x \) and \( y \), were related in such a way that when \( y=1 \) then \( x \) must equal 1, otherwise \( x \) can be either 0 or 1. In such a case we know that transition from \( xy=11 \) to \( xy=01 \) is not possible.

In our approach we utilize a fast logic decomposition to efficiently compute spatial correlation. The use of a fast decomposition in our approach makes it possible to speed up the computation of correlation by looking at a collection of small logic functions, rather than a few large ones. While many decompositions exist, [14] has been shown to be efficient in handling reasonably large logic functions.

To demonstrate our approach consider the example in Figure 2. It shows an example of a logic circuit, where the right-most AND gate \( f \) depends on two logic functions, \( g \) and \( h \), that have the same support. We now show how [14] is used to compute conditional probability of \( g \) given \( h \), allowing us to account for correlation of these two logic functions during toggle rate analysis of the AND gate producing signal \( f \) (encircled gate in Figure 2).

Figure 3 shows our initial setup, where logic functions for \( g \) and \( h \) are put side by side and variables \( a \) and \( b \) are assigned to the rows of the combined truth table. With this setup in place we apply the decomposition in [14] to find their common subfunctions. They are \( X_1=a \) and \( X_2=b \). The selector functions for \( X_1 \) and \( X_2 \) in function \( g \) are \( S_g=(cd) \) and \( S_g=d \). Given this decomposition, our goal is to determine the conditional probability \( P[g=1|h=1] \).

The basic idea of the following computation is to determine how each of the subfunctions, \( X_1 \) and \( X_2 \), that cause function \( g \) to assume a logic value 1 are affected by the logic state of function \( h \). To do so we first compute when
we can compute \( P[A|B] = \frac{P[A \cap B]}{P[B]} \)

doing so allows us to expand the probability of this expression into a sum of probabilities:

\[
\begin{align*}
P[A|B] &= \sum_{i} P[X_i S_i | h] - \sum_{j} P[X_j S_j | h] - \sum_{k} P[X_k S_k | h] - ...
\end{align*}
\]

A similar derivation is used to expand \( P[X_i S_i \cap h] \) terms with respect to function \( h \). When \( h \) is expanded, the resulting expression is a product of basis and selector functions for \( g \) and \( h \). Since basis functions have disjoint support from selector functions, the probability computation is simplified to the product of probabilities of two logic functions. One function has the support the same as the basis functions, while the other includes the support of the selector functions for both \( g \) and \( h \).

In the above derivation, functions \( X_i \) and \( S_i \) have disjoint support and therefore are assumed to be independent. While this assumption may not always hold and we will see errors as a results, to assume otherwise means using logic functions as a function of primary inputs of a circuit, which for all practical purposes is not feasible. However, as we will show through our results, the estimation error is acceptable.

\section{5. Experimental Results}

We conducted experiments to evaluate the effectiveness and processing time of the proposed approach on 14 MCNC circuits implemented on the Altera Stratix II FPGA. We used ModelSIM-Altera 6.0c and 10000 random vectors to generate the reference toggle rate for each benchmark circuit. We then ran the Quartus II 7.1 PowerPlay Power Analyzer to determine the actual toggle rate of each wire, accounting for inertial delay [1] that would cause some glitches not to
appear on a wire.

5.1 Experimental Procedure and Results
To generate results we used the transition density and static probability values from simulation to define the behaviour of the primary inputs of each circuit. We then applied our technique to estimate the toggle rate of each benchmark circuit and compared it to the results obtained via simulation.

We also applied the vectorless toggle rate estimation technique available in Quartus II 7.1 to the benchmark suite to see how our method fairs in comparison to one currently used in FPGA industry. The input to the Quartus II 7.1 PowerPlay Power Analyzer was the same as to our approach.

In both cases we compared the estimate to the actual toggle rate for each wire using the following formula:

\[
\text{%diff} = 100 \times \frac{(\text{estimated} - \text{actual})}{\text{MAX}(\text{estimated, actual})}
\]

We used the result for each wire in a circuit to determine the average percentage error as well as standard deviation for the error distribution. The results are shown in Table 1.

In Table 1 each circuit is listed by name in the left column. In column 2 we list the size in terms of ALUTs the circuit occupied on the Altera Stratix II FPGA. Columns 3 and 4 show the average error and standard deviation for toggle rate prediction made by Quartus II 7.1. In columns 5 and 6 we show the average error and standard deviation for our toggle rate prediction method. In column 7 we show the processing time needed for our approach to analyze toggle rate of each circuit. Finally, in column 8 we compare average error magnitude as computed by our method to Quartus II 7.1, while in column 9 we compare the standard deviation to that obtained by Quartus II 7.1.

5.2 Discussion
Our toggle rate estimation technique was able to achieve a 9.67% average error with a standard deviation of 23.52%. While there is room for improvement, the approach proved to be computationally efficient, where the processing time as a function of circuit size grows at a reasonable rate.

In most cases our approach was able to achieve a very good average error. It was particularly effective in cases where the circuit had a large number of inputs, where expressing LUT inputs in terms of primary circuit inputs was not a viable alternative. Examples of such circuits were C2670 and pair. Other circuits for which the average estimate error was low had few levels of LUTs and a low number of primary inputs. An example of such a case is alu4 and cps. For such circuits capturing the reconvergent fanout within 4 levels of ALUTs was sufficient to either incorporate primary inputs in the computation, or capture enough of the correlation to compute the toggle rate effectively.

This was not the case in three of our circuits, namely ex1010, pdc and spla. Each of those circuits has at most 16 primary inputs. However, what these circuits lack in the number of inputs they make up for in the number of logic levels required to compute the final result. As such, our approach of looking at 4 ALUT levels was insufficient to capture the spatial correlation, though exploring more levels would have lead to encompassing all primary inputs in the computation and producing a better result. Such approach, however, would not have worked well in circuits with larger number of inputs, as the processing time and the cone sizes would be large.

In addition to average error we are providing a standard deviation to show the consistency of our results. On average the standard deviation of the toggle rate prediction error is 23.52%. While this may seem large, the question becomes how significant this error is in terms of absolute estimate error. To determine this we analyzed absolute toggle rate error for the MCNC circuits to find out the magnitude of the error we can expect with our approach. The graph in Figure 5 shows absolute estimate error achieved by our approach.

### Table 1. Toggle Rate Estimation Results in Comparison to Quartus II 7.1.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>ALUTs</th>
<th>Quartus II 7.1 Average Error (%)</th>
<th>Quartus II 7.1 St. Dev.</th>
<th>Our Average Error (%)</th>
<th>Our St. Dev.</th>
<th>Our Processing Time (s)</th>
<th>Average Error Our vs. Quartus II 7.1 (%)</th>
<th>St. Dev. Our vs. Quartus II 7.1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>alu4</td>
<td>731</td>
<td>23.33</td>
<td>58.39</td>
<td>0.30</td>
<td>22.37</td>
<td>3</td>
<td>-98.71</td>
<td>-61.69</td>
</tr>
<tr>
<td>apex2</td>
<td>797</td>
<td>55.82</td>
<td>55.26</td>
<td>-2.03</td>
<td>29.49</td>
<td>2.9</td>
<td>-96.36</td>
<td>-46.63</td>
</tr>
<tr>
<td>apex4</td>
<td>645</td>
<td>22.01</td>
<td>74.74</td>
<td>-14.65</td>
<td>20.15</td>
<td>2.7</td>
<td>-33.44</td>
<td>-73.04</td>
</tr>
<tr>
<td>C2670</td>
<td>117</td>
<td>-8.84</td>
<td>22.56</td>
<td>-1.51</td>
<td>16.10</td>
<td>0.7</td>
<td>-82.92</td>
<td>-28.63</td>
</tr>
<tr>
<td>C3450</td>
<td>262</td>
<td>-8.04</td>
<td>45.56</td>
<td>-12.12</td>
<td>20.83</td>
<td>1.5</td>
<td>33.66</td>
<td>-54.28</td>
</tr>
<tr>
<td>cps</td>
<td>427</td>
<td>30.33</td>
<td>68.13</td>
<td>-0.14</td>
<td>40.26</td>
<td>1.4</td>
<td>-99.54</td>
<td>-40.91</td>
</tr>
<tr>
<td>dalu</td>
<td>193</td>
<td>20.31</td>
<td>53.98</td>
<td>-8.50</td>
<td>20.59</td>
<td>1.1</td>
<td>-58.15</td>
<td>-61.86</td>
</tr>
<tr>
<td>ex5p</td>
<td>504</td>
<td>10.25</td>
<td>76.24</td>
<td>-6.53</td>
<td>22.00</td>
<td>3.6</td>
<td>-36.29</td>
<td>-71.14</td>
</tr>
<tr>
<td>ex1010</td>
<td>1802</td>
<td>7.88</td>
<td>83.62</td>
<td>-25.47</td>
<td>21.73</td>
<td>31.2</td>
<td>69.06</td>
<td>-74.01</td>
</tr>
<tr>
<td>misex3</td>
<td>732</td>
<td>29.80</td>
<td>64.54</td>
<td>-8.06</td>
<td>22.09</td>
<td>2.8</td>
<td>-72.95</td>
<td>-65.77</td>
</tr>
<tr>
<td>pair</td>
<td>336</td>
<td>-2.07</td>
<td>32.10</td>
<td>-2.93</td>
<td>11.43</td>
<td>1.9</td>
<td>29.35</td>
<td>-64.39</td>
</tr>
<tr>
<td>pdc</td>
<td>1894</td>
<td>15.25</td>
<td>82.31</td>
<td>-26.57</td>
<td>25.12</td>
<td>30.7</td>
<td>42.60</td>
<td>-69.48</td>
</tr>
<tr>
<td>seq</td>
<td>854</td>
<td>44.55</td>
<td>60.92</td>
<td>-3.07</td>
<td>33.34</td>
<td>3.4</td>
<td>-93.11</td>
<td>-45.27</td>
</tr>
<tr>
<td>spla</td>
<td>1495</td>
<td>30.84</td>
<td>76.49</td>
<td>-24.13</td>
<td>23.80</td>
<td>21.4</td>
<td>-21.76</td>
<td>-68.88</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td>-9.67</td>
<td>23.52</td>
<td></td>
<td>-37.04</td>
<td>-59.00</td>
</tr>
</tbody>
</table>
The average error of the distribution shown in Figure 5 is 0.024 transitions per cycle and the standard deviation is 0.085 transitions per cycle.

5.3 Related Work

The work closest to ours is [7]. The work presented there uses BDDs to compute correlation between logic signals, and similarly to our approach create cones of logic up to a certain depth to capture spatial correlation. The main differences between the approach in [7] and ours is the use of logic decomposition to speedup toggle rate computation as well as the use of post-routing delay data to account for glitches.

In contrast to the work in [2], which also accounts for the presence of glitches, our approach uses the logic equation of a LUT to compute the transition density generated by a LUT and does not require calibration to work well with any given CAD tool. Instead, it depends only on device parameters, such as logic and interconnect delays, and inertial delay [1].

6. CONCLUSION AND FUTURE WORK

This paper presented a new approach to toggle rate computation that utilizes a fast XOR-based logic decomposition technique [14]. The decomposition technique was chosen because of its ability to quickly compute a functional decomposition of logic functions, even for logic cones that exceed 20 inputs. In contrast, other techniques would require more time to process cones of similar size, or the generated decomposition would consist of larger number of subfunctions than the technique presented in [14]. This helps us provide a better result in a shorter amount of time.

The method presented here was applied to circuits implemented on the Altera Stratix II FPGA, showing a significant improvement in the accuracy of toggle rate estimation as compared to the one currently used in the Quartus II commercial CAD tool.

A further area of study along the proposed approach is to further compress the data used in correlation computation. The main idea here would be to preserve only the logic nets that are part of the reconvergent paths, while the other nets could be compressed. This could speed up the correlation computation and allow the cone sizes to be much larger, thereby improving the accuracy of results.

7. ACKNOWLEDGMENTS

The authors would like to thank Dr. D. Lewis for feedback during the course of this work. We also appreciate the assistance of Dr. A. Grbic and B. Chan with running experiments using Quartus II software. We thank Altera Corporation for funding this research.

8. REFERENCES