Mitigating FPGA Interconnect Soft Errors by In-Place LUT Inversion

Naifeng Jing\textsuperscript{1}, Ju-Yueh Lee\textsuperscript{2}, Weifeng He\textsuperscript{1}, Zhigang Mao\textsuperscript{1}, Lei He\textsuperscript{2}

\textsuperscript{1} School of Microelectronics, Shanghai Jiao Tong University, Shanghai, China\textsuperscript{*}
\textsuperscript{2} Electrical Engineering Department, University of California, Los Angeles, USA

Abstract — Modern SRAM-based FPGAs (Field Programmable Gate Arrays) use multiplexer-based unidirectional routing, and SRAM configuration cells in these multiplexers contribute to the majority of soft errors in FPGAs. In this paper, we formulate an In-Place inVersion (IPV) on LUT (Look-Up Table) logic polarities to reduce the Soft Error Rate (SER) at chip level, and reveal a locality and NP-Hardness of the IPV problem. We then develop an exact algorithm based on the binary Integer Linear Programming (ILP) and a heuristic algorithm based on the Simulated Annealing (SA), both enabled by the locality. We report results for the 10 largest MCNC combinational benchmarks synthesized by ABC and then placed and routed by VPR. The results show that IPV obtains close to 4x chip level SER reduction on average and SA is highly effective by obtaining the same SER reduction as ILP does. A recent work IPD has the largest LUT level SER reduction of 2.7x in literature, but its chip level SER reduction is merely 7% due to the dominance of interconnects. In contrast, SA-based IPV obtains nearly 4x chip level SER reduction and runs 30x faster. Furthermore, combining IPV and IPD leads to a chip level SER reduction of 5.3x. This does not change placement and routing, and does not affect design closure. To the best of our knowledge, our work is the first in-depth study on SER reduction for modern multiplexer-based FPGA routing by in-place logic re-synthesis.

Keywords - SRAM-based FPGA; Soft Errors; SER; Routing; Interconnect; Fault Mitigation; Logic Re-synthesis; Logic Polarity

I. INTRODUCTION

Modern SRAM-based FPGAs use SRAM cells to configure logic and interconnects, numerically totaling up to 160 million in XilinxVirtex-6 \cite{1}. These FPGAs suffer from single event upset (SEU) induced soft errors, and their resilience against SEU decreases with technology scaling. Therefore, reducing the soft error rate (SER) for SRAM-based FPGAs has gained growing significance. Classic Triple Module Redundancy (TMR) employs circuit redundancy both in LUT and interconnect but with high overhead in area, power and performance. Recent logic re-synthesis techniques, such as ROSE \cite{2}, IPR \cite{3}, IPD \cite{4} and R2 \cite{5} reduce SER in LUTs by leveraging different logic masking techniques. Targeting on low or no cost in area, power, performance and design closure, they are preferred by non-mission-critical applications, e.g. networking and communications, which in fact are the primary applications of FPGAs. However, these techniques do not explicitly consider the interconnect SER and thus chip level SER reduction could be limited due to the interconnect dominates in FPGAs.

Modern FPGAs have shifted to multiplexer-based (MUX-based) unidirectional routing architecture \cite{6,7}, where the fault mechanism is different from conventional bidirectional routing in the previous studies \cite{8,9}. In this paper, considering MUX-based unidirectional routing, we formulate an In-Place inVersion (IPV) of LUT logic polarities to reduce the interconnect SER, and reveal a locality and NP-Hardness of the IPV problem. We then develop an exact algorithm based on the binary Integer Linear Programming (ILP) and a heuristic algorithm based on the Simulated Annealing (SA), both enabled by the locality. We report results for the 10 largest MCNC benchmark circuits mapped by ABC \cite{10} and then placed and routed by VPR \cite{11}. The results show that IPV can obtain nearly 4x reduction on chip level SER. In addition, the SA approach is highly effective, obtaining the same quality of results when compared to exact ILP solutions but is much faster in runtime. In contrast to the IPD algorithm with highest 7% chip-level SER reduction among \cite{2-5}, SA-based IPV obtains nearly 4x reduction and runs 30x faster. Furthermore, combining IPV and IPD leads to 5.3x SER reductions at the chip level. This does not change placement and routing, and thus has no impact on design closure. To the best of our knowledge, this work is the first in-depth study on SER reduction for modern MUX-based FPGA routing by the in-place logic re-synthesis.

The paper is organized as follows. Section II introduces the behavior and evaluation of the soft error on the unidirectional routing architecture. Next, we formulate the IPV problem in section III and present IPV properties and algorithms in section IV. Section V shows the experimental results and section VI concludes this paper.

II. PRELIMINARIES

A. FPGA Architecture and Interconnect Fault

An FPGA architecture is mainly defined by Configurable Logic Blocks (CLBs) and routing architectures as illustrated in Fig. 1. Interconnects are critical because they contribute a large portion of total FPGA area and total configuration bits. In our concerned unidirectional routing architecture, both the inter-CLB routing (including connection boxes and switch boxes) and intra-CLB routing employ directional MUXes to route signals. Each MUX is typically configured by several encoded configuration bits (CRAM bits), which contribute to the majority of the CRAM bits in an FPGA. For example, we observe that interconnects contribute to nearly 80% of the CRAM bits for the 10 largest MCNC benchmarks when they are synthesized to the minimum FPGA dimensions by 6-input LUT \cite{12}.

As illustrated in Fig. 1, when one of the bits flips its value due to a soft error on a MUX, an erroneous input signal (dotted) is then selected. If this signal from the faulty MUX reaches the...
primary outputs of the chip, a functional failure at the chip level occurs. Note that this fault mechanism is not a bridging fault as studied for bidirectional routing in [8][9].

B. Fault Rate Evaluation

We evaluate the failure rate for each CRAM bit under the single fault assumption\(^1\) by \(\text{SER}_b\), which is defined as follows.

**Definition 1:** For a circuit \(C\) with \(n\) primary inputs, the soft error rate of a CRAM bit \(b\), denoted as \(\text{SER}_b\), is the probability of functional failures on the circuit due to the SEU on bit \(b\).

\[
\text{SER}_b = \Pr \left( C_b(x) \neq C_b(x) | b \rightarrow \text{SEU} \rightarrow b \right) \tag{1}
\]

where \(x \in \{0,1\}^n\) is the exhaustive set of input vectors and \(C_b(x)\) is the circuit outputs under \(x\), and \(C_b(x)\) is the circuit outputs when \(b\) is flipped due to SEU. The \(\text{SER}_b\) can be obtained by exhausting \(2^n\) input vectors, which is a very time-consuming process. In practice, it can be approximated by a Monte Carlo based fault simulation of as many as \(K\) times, which can provide a good accuracy as studied in [14].

In general, the metric of \(\text{SER}_b\) applies to any circuit element as long as it is configured by CRAM bits. For example, we introduce \(\text{SER}_R\) in (2), as the total routing \(\text{SER}\) to evaluate the sensitivity of functional failure to all the CRAM bits in routing elements denoted by \(R\). We also quantify the circuit fault rate by all the CRAM bits in various elements in circuit \(C\) as in (3), which is referred as chip level \(\text{SER}\) in this paper.

\[
\text{SER}_R = \sum_{b \in R} \text{SER}_b \tag{2}
\]

\[
\text{SER} = \text{SER}_C = \sum_{b \in C} \text{SER}_b \tag{3}
\]

III. PROBLEM FORMULATION

A. Motivating Example of Fault Masking

As illustrated in Fig. 1, when MUX \(m\) has one of its CRAM bits \(b\) flipped due to SEU, the output is driven by \(j\) instead of the desired \(i\). If \(j\) carries a different logic value \(v(j)\) from \(v(i)\), a fault is injected onto the inputs of the immediate fan-outs of \(m\). However, if \(v(j)\) equals \(v(i)\), no fault is injected even if SEU happens. That is, the fault can be instantly masked at \(m\). In addition, \(\text{SER}_R\) also depends on the observability don’t care of MUX \(m\), \(obv(m)\), which indicates if the fault can be masked by logic during its propagation to circuit outputs as in [3]. As a result, \(\text{SER}_b\) can be given by \((v(i) \oplus v(j)) \cdot obv(m)\).

Figure 1. FPGA unidirectional routing and multiplexer structure.

Note that logic polarity can be independently determined on

---

\(^1\) Simultaneous multiple SEUs seldom happen in commercial FPGAs as reported in [13].
fan-in LUT pair (as stated by Property 2 below). Thus, our IPV problem tries to reassign the polarity for each L, so that \( \text{SER}_R \) can be minimized for all \( b_k \) \((k=1, \ldots, m)\).

IV. PROPERTIES AND ALGORITHMS

A. Properties of IPV Problem

**Property 1:** Our IPV problem is NP-Hard.

**Sketch of Proof:** IPV problem can be reduced from the binary Max-Sum (labeling) problem which is known to be NP-Hard [17]. We skip the details of the proof due to the limited space.

In IPV problem, when one or multiple LUTs are selected to be inverted for fault masking, \( \text{SER}_R \) should be updated accordingly after each inversion. Intuitively, each update needs a new pass of circuit fault simulation that is highly time-consuming. In this paper, we reveal that \( \text{SER}_R \) can be analytically updated by a pre-calculation of the bit fault rates based on property 2.

**Property 2** (Locality of Bit \( \text{SER} \) upon Polarity Inversion): Under the single fault assumption and for a given logic network, \( \text{SER}_R \) for a routing CRAM bit solely depends on the logic polarities of its pseudo fan-in LUT pair, independent of the polarities of the other LUTs in the network.

We also skip the proof due to the limited space here.

B. Locality based \( \text{SER} \) Calculation

Based on the locality property, the \( \text{SER}_R \) with possible inversions can be calculated for at most 4x complexity to \( \text{SER}_R \) as in (4).

\[
\text{SER}_R^4 = \begin{cases} 
\text{SER}_R^{00} + L(b) \& + l(b) \\
\text{SER}_R^{10} + L(b) \& - l(b) \\
\text{SER}_R^{01} - L(b) \& + l(b) \\
\text{SER}_R^{11} - L(b) \& - l(b)
\end{cases}
\]  

(4)

where the quadruplicate values of \{\text{SER}_R^{00}, \text{SER}_R^{01}, \text{SER}_R^{10}, \text{SER}_R^{11}\} are called a \( \text{SER} \) quadruplet of bit \( b \). Each quadruplet provides four error rates indicated by the two superscripted numbers, indicating if one of its pseudo fan-in LUTs is inverted or not. For abbreviation, we denote it as \( \text{SER}_R^4 \{P_{L(b)}, P_{l(b)}\} \), where \( P \) is a function indicating the polarities of LUTs \( L(b) \) and \( l(b) \), i.e., + or -. Thus, the total routing \( \text{SER}_R \) can be written as

\[
\text{SER}_R^4 = \sum_{b \in R} \text{SER}_R^4 \{P_{L(b)}, P_{l(b)}\} \]

(5)

Eq. (5) reveals that \( \text{SER}_R \) for a given circuit can be updated as an algebraic sum on each CRAM bit by its \( \text{SER} \) quadruplet. To get their values, we currently use the fault simulation method that is within 4x complexity of the \( \text{SER}_R \) in (2). In this way, the iterative fault simulation after each reassignment of LUT polarity can be avoided.

C. Binary ILP Based Algorithm

In this section, we use a binary ILP formulation to provide an insight on the capability of IPV improvement. We take a set of binary variables \( x_i \) to denote whether a LUT \( i \) is inverted or not, i.e., a positive LUT has its \( x_i \) as 0. A binary inverting quadruplet \([f_{b}^{00}, f_{b}^{01}, f_{b}^{10}, f_{b}^{11}]\) is used to denote the polarities of the pseudo fan-in LUTs for each routing bit \( b \). As a result, the binary ILP formulation for our IPV problem is given by

\[
\text{min } \text{SER}_R^4 = \sum_{b \in R} \text{SER}_R^4 \{f_{b}^{00}, f_{b}^{01}, f_{b}^{10}, f_{b}^{11}\} \\
\text{s.t. } f_{b}^{00} \leq 1 - x_i, \\
f_{b}^{01} \leq 1 - x_i, \\
f_{b}^{10} + 1 \geq 1 - x_i + 1 - x_i,
\]

(6)

where \( x_i = L(b), x_i = l(b) \). This set of constraints models the fact that exactly one value in the quadruplet \{\text{SER}_R^{ij}\} should be selected for each bit \( b \) using \( f_{b}^{ij} \) as a mask. Other constraints on \( f_{b}^{00}, f_{b}^{01}, f_{b}^{10} \) and \( f_{b}^{11} \) can be similarly written as in (7).

By forcing the corresponding \( x_i \) values to be 0 in the constraints, our ILP formulation also applies to the situation where some LUT input or output polarities are not invertible.

D. Simulated Annealing Based Algorithm

We also propose a Simulated Annealing (SA) based algorithm to solve IPV efficiently while providing good quality of routing \( \text{SER} \) reduction compared to ILP. The SA based algorithm starts from the initial circuit with positive logic polarities for all the LUTs. Then, it switches to another LUT polarity assignment by inverting a random LUT polarity at each move. The objective function of the new assignment is evaluated by (5). New assignment with a better cost is always accepted while the worse assignment is accepted conditionally based on the acceptance probability. The annealing starts from a temperature of 0.008, and is updated by a decreasing factor of 1.003. It continues till the minimum temperature of 2.0e-6 is reached.

E. Overall Algorithm Flow

As illustrated in Fig. 6, our IPV algorithm mitigating SEU fault on FPGA interconnects by LUT logic polarity inversion consists of three phases. Starting from the given netlist of a circuit, it first applies logic optimization and technology mapping. The mapped circuit is packed into logic blocks, and then placed and routed by physical design tools. Secondly, in order to obtain the bit \( \text{SER} \) values, we develop an SEU fault analysis framework which starts right after P&R. It performs logic simulation based on post-layout circuit information to calculate
Atomic logic inversion

Logic Optimization & Technology Mapping

Synthesis and Placement & Routing

Architecture File

Criticality quadruplet

ILP model & solver

Simulated Annealing

Figure 6. The overall flow of our IPV algorithm

the fault rate for each CRAM bit in routing MUXes in the form of SER quadruplet. This is the basis for both the binary ILP and the SA-based approaches. Finally, we start the ILP solver and the SA-based approach to seek the optimal reassignment of the polarities for all the LUTs. The result with maximum inter- connect fault rate reduction is then selected and back-annotated to the initial circuit by the atomic logic inversion operations to finish our proposed re-synthesis flow.

V. EXPERIMENTAL RESULTS

A. Experimental Settings

In this section, we report the experiments for the 10 largest MCNC combinational circuits as in [4]. We used parameterized architecture in VPR [11] to characterize different FPGA architectures. Firstly, we perform logic optimization and technology mapping onto 4 and 6-input LUTs by Berkeley ABC [10]. Each mapped circuit is packed by two different CLB architectural settings respectively, i.e. 4-input LUT with cluster size 4 and 6-input LUT with cluster of 8. Then, VPR was used to implement a minimum dimension FPGA array without incurring extra unused CRAM bits that exceed the actual need of the circuit. We applied a Monte Carlo based fault simulation to generate the bit SER quadruplets. Note that our IPV algorithm applies to either combinational or sequential circuits to mitigate interconnect SEU faults as long as the bit SER quadruplets are available. We used Mosel as our ILP solver to seek the globally optimal assignment of the LUT polarities.

B. Comparison between the ILP and SA approaches

Table I first presents size statistics of the benchmark circuits. Next, it compares fault rate reductions in terms of SER ratios before and after applying IPV by ILP and SA approaches for all routing MUXes. From the table, one can see that both ILP and SA approaches significantly reduce SER. For example, for a 4-input LUT with cluster size 4, the interconnect SER is reduced by 1.2x to 1.72x with an average around 6x. For a 6-input LUT with cluster size 8, the SER is reduced about 5.4x on average. In addition, by considering the CRAM bit percentage, the SER can be reduced by 3.97x and 3.67x on average at the chip level for the 4-input and 6-input LUTs, respectively.

The SER reduction for 6-input LUT and 8 LUTs per cluster is slightly smaller, because larger LUT and cluster sizes have fewer interconnects and fewer MUXes that limit the room for improvement. While it is natural for different circuits to obtain different improvements, “des” has the lowest and much smaller SER reduction. This is because the SER quadruplicate values are high and close to each other in “des”, which limits the design freedom that can be leveraged by IPV.

Table I also reports runtimes. The runtime excludes the fault simulation time for SER quadruplets, which is relatively small compared to the runtime consumed by ILP. From the table, one can see that ILP is able to solve most of the circuits optimally while SA can obtain the same SER reductions as ILP but runs almost 100x faster. For the other circuits as marked, where a timeout of 10 hours is applied to the ILP solver like in “des”, SA obtains slightly higher SER reductions. These results show that the SA approach is highly effective and efficient for IPV problem. It is worthwhile to point out that circuits like “des” are not the largest circuits in experiments, so the efficiency of ILP depends on both circuit size and structure. When ILP and SA obtain the same SER reductions, LUT inversions in their solutions are often not the same and ILP inverts fewer LUTs in general. This implies that there are multiple “optimal” solutions from the point of view of SER reduction.

<table>
<thead>
<tr>
<th>Circuit</th>
<th>LUT size k=4, Cluster size N=4</th>
<th>LUT size k=6, Cluster size N=8</th>
</tr>
</thead>
<tbody>
<tr>
<td>#LUT</td>
<td>Int. SER Reduce</td>
<td>Int. CRAM bit%</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>ex5p</td>
<td>622</td>
<td>2.51</td>
</tr>
<tr>
<td>alu4</td>
<td>744</td>
<td>2.04</td>
</tr>
<tr>
<td>misc3</td>
<td>773</td>
<td>3.05</td>
</tr>
<tr>
<td>apex4</td>
<td>821</td>
<td>4.79</td>
</tr>
<tr>
<td>apex2</td>
<td>1014</td>
<td>3.91</td>
</tr>
<tr>
<td>seq</td>
<td>1084</td>
<td>3.32</td>
</tr>
<tr>
<td>ex1010</td>
<td>1120</td>
<td>7.26</td>
</tr>
<tr>
<td>des</td>
<td>1750</td>
<td>1.10</td>
</tr>
<tr>
<td>spla</td>
<td>2229</td>
<td>17.22</td>
</tr>
<tr>
<td>pdc</td>
<td>2304</td>
<td>14.60</td>
</tr>
<tr>
<td>Avg.</td>
<td>–</td>
<td>5.99</td>
</tr>
</tbody>
</table>

TABLE I. INTERCONNECT AND CHIP LEVEL SER REDUCTION FOR 10 BENCHMARK CIRCUITS
TABLE II. CHIP LEVEL SER REDUCTION COMPARED TO PREVIOUS WORK

<table>
<thead>
<tr>
<th>Circuit</th>
<th>IPV SER Reduce</th>
<th>IPD SER Reduce</th>
<th>IPD+IPV Chip SER Reduce</th>
<th>Runtime (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Int. Chip LUT</td>
<td>Chip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ex5p</td>
<td>1.90 1.78</td>
<td>3.60 1.04</td>
<td>1.97</td>
<td>26 795</td>
</tr>
<tr>
<td>acl4</td>
<td>1.99 1.82</td>
<td>3.88 1.09</td>
<td>2.09</td>
<td>27 1466</td>
</tr>
<tr>
<td>nixes3</td>
<td>2.98 2.50</td>
<td>5.58 1.09</td>
<td>3.12</td>
<td>29 1235</td>
</tr>
<tr>
<td>apex4</td>
<td>4.91 3.86</td>
<td>4.59 1.09</td>
<td>4.89</td>
<td>43 1430</td>
</tr>
<tr>
<td>apex2</td>
<td>3.75 2.82</td>
<td>8.39 1.12</td>
<td>4.02</td>
<td>51 1137</td>
</tr>
<tr>
<td>seq</td>
<td>3.40 2.78</td>
<td>5.72 1.09</td>
<td>3.53</td>
<td>58 1659</td>
</tr>
<tr>
<td>evx1010</td>
<td>7.46 4.94</td>
<td>4.27 1.07</td>
<td>7.05</td>
<td>56 1635</td>
</tr>
<tr>
<td>des</td>
<td>1.09 1.09</td>
<td>1.48 1.02</td>
<td>1.10</td>
<td>34 2022</td>
</tr>
<tr>
<td>spla</td>
<td>14.05 7.77</td>
<td>7.04 1.07</td>
<td>13.23</td>
<td>142 3270</td>
</tr>
<tr>
<td>rde</td>
<td>12.51 7.31</td>
<td>8.63 1.07</td>
<td>12.17</td>
<td>154 3429</td>
</tr>
<tr>
<td>Avg.</td>
<td>5.40 3.67</td>
<td>5.32 1.07</td>
<td>5.32</td>
<td>62 1808</td>
</tr>
</tbody>
</table>

C. Chip Level SER Reduction Compared to Previous Work

Because SA is effective and efficient, SA is used for IPV in the rest of this paper. Table II reports chip level SER reductions, by comparing IPV with a previous IPD algorithm [4] on mapped 6-input LUTs. Although IPD significantly reduces the fault rate on LUTs (around 5x), its chip level SER reduction is merely 7%, because it does not consider interconnects explicitly. Therefore, future SER mitigation algorithms should be developed with explicit consideration of interconnects.

In Table II, we also evaluated the combined algorithm of IPD+IPV on 6-input LUTs, where the IPV is applied after IPD. Because IPV is performed within each LUT, it is orthogonal to IPV. From the table, we see that the combination of these two algorithms reduces chip level SER by 5.3x on average. With much reduced interconnect SER by IPV, the SER reduction by IPD for LUTs (or in general, by algorithms from [2-5]) gains more significance. Finally, Table II compares runtimes for IPV and IPD and shows that IPV runs 30x faster on average.

VI. CONCLUSION AND FUTURE WORK

By targeting on routing multiplexers that are the dominant routing elements in the modern unidirectional FPGA routing architecture, we have identified a new fault masking mechanism and formulated an IPV problem for In-Place inVersion of LUT logic polarities to maximize fault masking for all interconnect multiplexers. We have shown that the problem is NP-Hard and also revealed a locality for IPV. Due to the locality, we have developed two algorithms based on ILP and SA approaches. Experiments have shown that IPV obtains nearly 4x reduction on average for the chip level SER. In addition, SA is effective and efficient because it obtains the same fault reductions as ILP, if not better, but runs almost 100x faster. Note that IPV does not change placement and routing, therefore it is an in-place optimization. It should have little or no impact on design closure. This will be verified in the future.

IPV has obtained close to 4x reduction on average for the chip level SER. The best existing in-place algorithm is IPD [4], which has merely about 7% chip-level SER reduction but runs 30x slower. However, combining IPV and IPD leads to a chip-level SER reduction of 5.3x on average. This is because IPV and IPD target at interconnects and LUTs respectively and they are complementary to each other.

While the proposed IPV algorithms and implementations apply to both combinational and sequential circuits, we have not yet developed SER calculation for sequential feedbacks. This will be the focus of our future work. Furthermore, the interaction between IPV and those techniques from literatures such as [2-5] will be studied for further SER mitigation.

REFERENCES