An Analytical Model for Hardened Latch Selection and Exploration

Michael Sullivan, Brian Zimmer, Siva Hari, Timothy Tsai, Stephen W. Keckler
{misullivan, bzimmer, shari, timothyt, skeckler}@nvidia.com
NVIDIA

Abstract—Hardened flip-flops and latches are designed to be resilient to soft errors, maintaining high system reliability in the presence of energetic radiation. The wealth of different hardened designs (with varying protection levels) and the probabilistic nature of reliability complicates the choice of which hardened storage element to substitute where. This paper develops an analytical model for hardened latch and flip-flop design space exploration. It is shown that the best hardened design depends strongly on the target protection level and the chip that is being protected. Also, the use of multiple complementary hardened cells can combine the relative advantages of each design, garnering significant efficiency improvements in many situations.

Index Terms—Latch, Flip-Flop, Radiation Hardening by Design, Analytical Model, DICE, RCC, SEUT, Biser, Quatro.

I. Introduction

Soft errors in computer storage continue to be a major concern for mission-critical, high-availability, and high performance systems. Large on-chip storage arrays and transmission busses can be effectively protected by error correcting codes (ECC) [1]. Strong protection must be holistic in nature, however, and it is likely that unstructured storage elements which cannot be easily protected by ECC—such as the latches and flip-flops in the control path and random logic—will eventually degrade system reliability if left unprotected [2].

Radiation hardened storage elements are an attractive tool for protecting such unstructured latches and flip-flops. These hardened elements are more resilient to soft errors and they provide a mechanism to achieve a desired on-chip failure rate with minimal design changes. Instead of requiring micro-architectural changes, the vulnerable elements are simply swapped out with larger but more resilient versions of the same cell.

While the application of hardened storage elements requires only minimal changes to the chip, their effective use comes with hidden design complexity. There are a large number of potential hardened latch and flip-flop designs with significantly different levels of protection and expected overheads [3], [4], [5], [6], [7], [8], complicating design space exploration. In addition, architectural and application-level masking cause individual storage elements to differ in their contributions to the overall system failure rate [9], [10], [11].

This paper presents an analytical model to serve as a tool for selecting which and how many hardened storage elements a designer should insert to achieve the desired on-chip failure rate while minimizing overheads. Specifically, this paper makes the following contributions. It:

- Presents a simple yet useful analytical model that expresses the cost and FIT reduction following the intelligent selective insertion of hardened latches. This model closely matches the empirical data that are available from this tradeoff space.
- Demonstrates by use of this model that no one hardened latch design is best for all applications—the optimal hardened latch depends on the processor design as well as the FIT reduction target (or, conversely, the area and power constraints of the system).
- Considers and analyzes the use of multiple complementary hardened latches as appropriate. An example use of two or three complementary hardened latches reduces the estimated area and power overheads relative to the best single hardened latch by 30–60% in large parts of the design space.

The paper proceeds as follows. We first describe the different approaches to harden and selectively protect latches in Section II. Section III presents an analytical model that is able to characterize the effectiveness of selective latch insertion. Section IV evaluates existing hardened latches on some plausible systems, finding that the best hardened latch design depends on the system. Section V considers the use of multiple complementary hardened latches for superior efficiency. Finally, Section VII discusses some extensions of this work and concludes the paper.

II. Background

This section summarizes the background necessary for a complete understanding of the system-level hardened latch design space. A note on terminology in this paper: the terms latch, flip-flop, and sequential element are used interchangeably. Hardening procedures can be applied to a variety of latch designs; the model in this paper applies regardless of the type of sequential elements used in a chip.

A. Latch Hardening

Radiation hardened latches and flip-flops are designed to be resilient to particle-induced soft errors. There are three prevalent approaches towards hardening a sequential element. Strike Suppression techniques tailor the design and...
layout of a latch to increase the critical charge that is required to cause an error. Some examples of this type of latch protection include RCC [6] and LEAP [7]. The effectiveness of strike suppression techniques is limited because the critical charge cannot be raised enough to entirely prevent particle events from causing errors. However, the overhead for these techniques is also low, minimizing their barrier to adoption and making them an attractive choice from the system level.

**Redundant Node** storage elements use two interleaved nodes to store data and rely on internal feedback to restore the correct data in the presence of an error in one of these two nodes. Examples of this type of hardened latch include DICE [3], BISER [4], Quatro [5], [12], and SEUT [6]. Redundant node approaches are more effective than strike suppression, because particle strikes that exceed the critical charge will not cause an error at the output of the cell. However, the overhead of redundant node designs is also much higher than strike suppression techniques because most devices in the latch need to be duplicated to form the redundant structure. In-situ swapping of vulnerable cells with redundant node cells can therefore have a major impact on the overall design, increasing their barrier to adoption. Also, redundant node approaches are still vulnerable to errors when charge is collected by multiple nodes, so the redundant node must be physically separated as much as possible to maintain maximum resiliency, further increasing area overheads [6]. Strike suppression and redundant node techniques have also been applied together, most notably in the LEAPDICE [8] cell.

Finally, **Triple Modular Redundancy** (TMR), where the same bit is stored in three locations and a majority voter is used to correct a bit flip in any location, can be viewed as a hardened flip-flop. TMR cells have extremely high resiliency, as the constituent latches are physically isolated enough to guarantee that two independent events are needed to cause an error. However, their overhead is very large, as three flip-flops are needed in place of one. While the exact implementation details, resiliency improvements, and area overheads of hardened latches are still an open area of research, these three general approaches—strike suppression, redundant node, and TMR—serve as a useful categorization that represents both current and future techniques. In any case, there is a variety of hardened latch designs that create a rich tradeoff space. The analytical model in this paper can serve to aid the designer in selecting which hardened design (or designs) best suit a particular chip and reliability target.

### B. Selective Hardened Latch Insertion

To efficiently utilize hardened storage cells, a designer must be able to decide which sequential elements to protect. Prior studies have shown that processors exhibit a non-uniform architectural vulnerability factor (AVF) [13] across their latches—some storage elements are much more sensitive to transient errors than others. In addition, prior work has demonstrated that it is possible to characterize this asymmetric sensitivity and selectively protect the most crucial elements without incurring the high overheads of hardening every latch and flip-flop.

Figure 1 shows the asymmetric sequential element sensitivity observed in four prior studies. The horizontal axis of each sub-plot shows the fraction of latches chosen for (perfect) protection, assuming that the most critical latches are protected first; the vertical axis shows the relative FIT reduction among all sequential elements. The data are extracted from the figures of these papers using a computer-aided extraction tool [14]. (The fitted model from Section III is also overlayed on top of this empirical data, showing the fit of the analytical model proposed in this paper.) Ebrahimi et al. [11] use an iROC reliability characterization flow to analyze an OR1200 embedded processor [15]. This is the most relevant data for this paper, since it represents the full-chip analysis of a complete—albeit simple—processor. Hill et al. [9] analyze the sensitivity of a fixed and floating-point multiplier, which have a low and high AVF bias, respectively. Holcomb et al. [10] analyze the sensitivity of a small CMP router chip.

While the precise reliability characterization methodology of different prior studies differs, their general goal is the same: to rank the relative sensitivity of each sequential element. As an explanatory example, the general steps of characterization methodology from [11] follow:

1. Element characterization (to extract raw FIT rates)
2. Masking analysis
3. Pruning of the search space
4. Error injection with workloads

Steps (1) and (2) of this flow estimate the raw error rate of each sequential element, taking into account differences between the different cells (which can be significant [16]) and the temporal masking of latches that are on or near the critical path [17]. Step (3) is crucial for scaling this
methodology to large chips with many sequential elements. Finally, statistical fault injection is performed in step (4) with a diverse set of representative workloads.

During initial design space exploration, the error sensitivity asymmetry does not need to be precisely characterized. However, a reasonable approximation of this asymmetry enables many interesting conclusions about the overall effectiveness of latch hardening in reducing the system failure rate. The model in this paper makes extensive use of the characterized asymmetry of a chip, and shows it to be an important factor when making effective system-level decisions.

III. AN ANALYTICAL MODEL FOR HARDENED LATCH DESIGN SPACE EXPLORATION

As Section II shows, prior efforts have investigated a wide variety of hardened latches and other works have considered the selective protection of sequential elements. These efforts have largely progressed independent of one another—hardened design papers do not consider the system-level impact of the latches, and selective insertion papers assume that hardening provides perfect protection against soft errors. This paper proposes an analytical model to aid in the high-level design space exploration of imperfect hardened latches that are selectively inserted to achieve a target FIT rate. This model incorporates the important aspects that determine the system-level overheads and the level of protection that are afforded by selective latch hardening, with the goal of helping a designer to achieve a desired failure rate while keeping resilience overheads to a minimum.

The first step for this explanatory model is to fit a family of curves to the data that a designer receives from a latch reliability characterization flow such as those described in Section II-B. While the asymmetry of sequential element sensitivity depends on the underlying chip architecture, two properties of selective hardened latch insertion are clear:

- With no hardened elements, the relative FIT reduction is zero. With all elements hardened the relative FIT reduction is determined by the strength of the hardening technique.
- Given effective hardened latch selection, the rate of FIT reduction with each hardened latch monotonically decreases.

Equation 1 gives a latch AVF bias curve (ABC) function that captures the above behavior and shows compelling agreement with the empirical hardened latch selection data from Figure 1. The model takes three inputs and returns the relative FIT reduction expected across all sequential elements. Its input parameters are: (1) \( \text{RED} \), the relative SEU reduction provided by the hardening technique (10x, 20x, 100x, etc., higher is more protected), (2) \( \beta \), a parameter representing the asymmetry of storage elements’ error sensitivity (\( \beta > 0 \), higher is more asymmetric), and (3) \( \text{HFRAC} \), the fraction of all storage elements that are selectively hardened (0 \( \leq \text{HFRAC} \leq 1 \), higher is more protected).

The ABC curve is shown in Figure 2. A variety of parameterized curves representing imperfect hardened latches. Latches with a RED of 2x, 4x, 8x, and 10000x are shown; each curve is labeled appropriately. Curves are shown for a chip with an AVF bias of \( \beta = 5.0 \).

\[
\text{ABC(RED, } \beta, \text{ HFRAC)} = \begin{cases} 
1 - \frac{1}{\text{RED}} & \text{if } \text{HFRAC} = 1 \\
\frac{1 - e^{-\beta \text{HFRAC}}}{1 - e^{-\beta}} & \text{otherwise}
\end{cases}
\]

Figure 2 illustrates how the ABC curve captures the behavior of imperfect latch protection. The ABC curve is shown for hardened latches with a RED of 2x, 4x, 8x, and 10000x. Even with all latches in a design hardened (HFRAC = 1.0) the overall sequential FIT reduction cannot exceed the protection level afforded to each individual element.

Figure 3 demonstrates the second component of the AVF bias curve: the impact of asymmetric latch sensitivity. The \( \beta \) input to the ABC curve characterizes this asymmetry—a high \( \beta \) indicates that a small number of sequential elements dictate the overall soft error rate. Non-linear least-squares regression is used to fit a \( \beta \) to the four chips or circuits in Figure 1. (This \( \beta \) minimizes the mean squared error with the observed sensitivity data.) The uniform sensitivity curve (\( \beta \) approaching 0) is also shown; it can be seen that a uniform sensitivity assumption is inappropriate for even the least biased system (the fixed-point multiplier from [9], with \( \beta = 4.57 \)).

A. Extracting Overhead Estimates

The fraction of latches that need to be protected to achieve a target reduction in latch failure rate, \( t_{\text{FIT}} \) (0 \( \leq t_{\text{FIT}} \leq 1 \), higher is more protected), can be found by solving \( \text{ABC(RED, } \beta, \text{ HFRAC)} = t_{\text{FIT}} \) for HFRAC. The symbolic solution follows.

\[
\ln \left( \frac{e^{\beta (1 - \text{RED})}}{e^{\beta (t_{\text{FIT}} - 1)} - 1 - e^{\beta}} \right) - \frac{t_{\text{FIT}}}{\beta} \to \text{HFRAC}
\]

Given a fraction of protected latches from Equation 2, it is straightforward to estimate the area and power overheads of protection directly from the overheads of the selected hardening technique.
The necessary changes to the model from Section III follow. Taking a set of $N$ different Pareto-optimal hardening techniques (such as those from Table I), sort the techniques from the strongest to the weakest—$eta_{\text{red}}$—and with it comes additional area overheads if the target hardening becomes cost-effective in different points in the design space. This section extends the model to consider the simultaneous application of multiple hardened latches.

V. MULTIPLE COMPLEMENTARY LATCH EVALUATION

Figure 4b shows the fraction of latches that need to be protected to achieve different FIT reductions with an SEUT cell ($\beta=37x$). The behavior is similar to that of perfect hardening, so long as the target FIT reduction is much less than $37x$. As the target approaches $37x$, however, there is an exponential increase in the fraction of latches that must be protected. This is true for all values of $\beta$, but those systems with a high AVF bias are able to get much closer to the limit before entering this costly region.

It is apparent from Figure 4 that the most efficient hardened cell depends on the latch AVF bias and the FIT reduction required across all sequential elements. Figure 5 shows this more clearly by plotting the area overhead of the three latches from Table I across systems with a low, medium, and high AVF biases. Several findings are clear. The impact of chip AVF on expected overheads can be seen—it is always less costly to protect a highly biased chip, and the difference between the low ($\beta=5$) and medium ($\beta=15$) biased chips is striking. All protection techniques demonstrate the attractive property of exponential increases in FIT reduction for approximately linear cost increases so long as the target FIT reduction is much less than the RED of the hardening technique. As the target FIT reduction approaches RED, however, there is an exponential cost explosion and eventually another protection technique should be preferred.

It can be seen from Table I that the strength of the hardened latches (apart from TMR) suffers from a large degree of uncertainty. Many factors can impact this uncertainty, including the unpredictability of future technology generations [19], [18], [20], measurement variability and testing factors [21], and temperature dependence [22]. The worst-case conditions for SEUT are shown in Figure 5 by a dashed line. The shape of the overhead curve does not change, but the maximum FIT reduction of each approach decreases (and with it comes additional area overheads if the target FIT reduction is near the limits of this weakened capacity).

The results of the model apply only to flip-flop protection and need to be interpreted from a holistic point of view. If flip-flop failures contribute only 25% of the total system FIT, then the difference in system FIT for a $10x$ versus $100x$ reduction in flip-flop FIT is actually very small. In addition, if flip-flops consume 20% of chip area, even a 30% hardening overhead translates to a system area overhead of only 6%.

IV. DESIGN SPACE EXPLORATION USING THE MODEL

There are a wide variety of hardened latches available to the designer that vary in their level of protection and overheads. For simplicity, we limit our evaluation to model a single hardened latch from each of the three hardened design classes in Section II-A. We choose a strike suppression latch (RCC) and a redundant node latch (SEUT), each characterized by Intel in a 22nm production library [18], and also estimate a TMR hardened latch. Table I gives the parameters of these hardened cells.

The asymmetry of on-chip latch sensitivity, characterized by $\beta$, tends to reduce the fraction of elements that need be hardened—the higher $\beta$ is, the fewer latches need be protected for a given overall FIT reduction. Figure 4 illustrates this through several contour plots; each curve on the plot represents a 5% increase in the number of hardened latches that are required to reach a target flip-flop FIT reduction. Figure 4a presents the data for an extremely strong latch hardening technique, and Figure 4b and shows an SEUT [18] cell with 37x SEU reduction.

It can be seen from Figure 4a that for low target FIT reduction levels a relatively small number of latches or flip-flops need be hardened. This agrees with prior findings [11], [9], [10], and it is true even for lower AVF biases—at $\beta=5$, for instance, only 13.7% of elements need be hardened to achieve a 2x FIT reduction. Conversely, it becomes increasingly expensive to provide high levels of FIT reduction, though chips with high AVF biases ($\beta>15$) can potentially achieve >99% FIT reduction with less than 30% of storage elements hardened.

Figure 3. The chip AVF bias $\beta$ for curves that are fit to the data from Figure 1. The uniform sensitivity curve ($\beta$ approaching 0) is also shown.

TABLE I

<table>
<thead>
<tr>
<th>Latch Type</th>
<th>Area Overhead</th>
<th>SEU Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1x</td>
<td>1x</td>
</tr>
<tr>
<td>Strike Suppression (RCC)</td>
<td>1.15x</td>
<td>6.3x ± 1.9x</td>
</tr>
<tr>
<td>Redundant Node (SEUT)</td>
<td>2x</td>
<td>37x ± 23x</td>
</tr>
<tr>
<td>Triplicated (TMR)</td>
<td>3.5x</td>
<td>1,000,000x</td>
</tr>
</tbody>
</table>

The hardened latches used in the design space exploration.
as before, partition \( hfrac \) among these \( N \) hardened latches using the stronger hardened designs for the more critical sequentials. This partition can be expressed by Equations 3 and 4, where \( hfrac_n \) is the fraction of latches protected by the hardening technique with protection level \( \text{RED}_n \) and \( hfrac_{1:n} \) is the cumulative sum of \( hfrac_n \) up to \( hfrac_{1:N} \) (and \( hfrac_{1:0} = 0 \)).

\[
\text{hfrac}_{1:M} = \sum_{n=1}^{M} hfrac_n \tag{3}
\]

\[
\text{hfrac} = \text{hfrac}_{1:N} \tag{4}
\]

The overall relative FIT reduction, analogous to \( \text{ABC(\text{RED}, \beta, \text{hfrac})} \) in the single hardened latch scenario, can be expressed through \( \text{ABC}_N(\text{RED}_n, \beta, \text{hfrac}_n) \) in Equation 5 where \( \text{RED}_n \) and \( \text{hfrac}_n \) are both vectors of length \( N \).

\[
\text{ABC}_N(\text{RED}_n, \beta, \text{hfrac}_n) = \sum_{n=1}^{N} \text{ABC}(\text{RED}_n, \beta, \text{hfrac}_1{(n-1)}) \tag{5}
\]

Figure 6 illustrates how multiple hardening techniques can combine for superior efficiency by way of a synthetic example. The \( \text{ABC}_N \) curve is shown for a system combining three theoretical hardening techniques: (A) \( \text{RED=8x SEU reduction at a 3.5x area overhead} \), (B) \( \text{RED=4x at a 2.5x overhead} \), and (C) \( \text{RED=2x at a 1.5x overhead} \). The multi-colored solid line shows \( \text{ABC}_N \) for a heterogeneous scheme that protects 15% of the system latches with design (A), 15% with (B), and 30% with design (C). The three dotted lines represent the corresponding \( \text{ABC} \) curves if each hardened latch were used separately. The relative area overhead of each approach is labeled on the right. It can be seen that the heterogeneous approach is the most efficient scheme—it incurs only a 75% overhead, significantly less than latch (B) at 90% despite providing better FIT reduction. The combined scheme has an even more significant cost benefit over design (A) with an equivalent level of protection (92% overhead, highlighted by an arrow in the figure). Note that a system with low AVF bias (\( \beta = 5 \)) is shown to increase the visibility of the FIT reduction difference between design (A) and the combined approach. This is a conservative scenario—the combination of multiple hardening techniques works better with more asymmetric latch sensitivity.

Figure 7 gives the overhead curves for the three hardened latches from Table I along with the four
hardening schemes. RCC+SEUT+TMR has an overhead roughly equivalent to that of RCC up to a FIT reduction of 6.3x. From that point onwards it performs roughly like SEUT+TMR until the FIT reduction approaches 37x, after which it provides a 44% improvement over TMR alone.

VI. Discussion

A. Overlap with Alternative Protection Schemes

There are higher-level protection schemes that are appropriate for certain on-chip structures; some, such as residue checking for large parallel multipliers [23], [24], might be significantly stronger than latch hardening at the same overhead. These alternative protection mechanisms should be preferred over latch hardening if they are more efficient and if their design effort is not prohibitive. The analytical model in this paper aids the design space exploration in the presence of these alternative mechanisms in two ways. First, it can compare the cost of latch hardening with alternative protection schemes to judge which is more efficient. Second, even with parts of the chip protected via specialized protection mechanisms, some unprotected flip-flops and latches will remain. Latch hardening can be seen as a catch-all technique for these leftover storage elements. This treatment of latch hardening as a complement to error coding can be seen in successful error protection efforts [2].

B. The Timing Overheads of Latch Hardening

While hardened latches incur some timing overheads, the analytical model in this paper does not consider the impact of this delay on system efficiency. One attractive possibility is to avoid hardened latch insertion wherever it would impact the clock period—a similar opportunistic insertion policy called slack-based flip-flop assignment has been employed in the past [25]. It is interesting to note that temporal masking will tend to lessen the sensitivity of latches that are on or near the critical path [17], such that the overall system-level impact of this optimization may be minimal. An accurate latch sensitivity characterization flow, such as the one from [11], is aware of temporal masking and it should be possible to incorporate a slack-based constraint into the latch selection procedure. Such extensions, and the timing component of the analytical model in this paper, are left for future work.

VII. Conclusion

This paper proposes and demonstrates the use of a model that allows the designer to quickly and transparently explore the tradeoff space of hardened latch designs and selective insertion. This model is used to demonstrate concretely that no single hardened latch design is optimal for all systems. An extension to the model also allows for the novel exploration of multiple complementary hardened latch designs; the use of multiple hardening schemes is able to reduce the overhead relative to the best single hardened latch by 30–60% in large parts of the design space.
References


