Static Voltage Over-scaling and Dynamic Voltage Variation Tolerance with Replica Circuits and Time Redundancy in Reconfigurable Devices

Dawood Alnajjar, Masanori Hashimoto, Takao Onoye Dept. of Information Systems Engineering, Osaka University & JST, CREST Osaka, Japan Email: dawood.alnajjar, hasimoto, onoye@ist.osaka-u.ac.jp

Abstract—This paper studies performance and timing failure probability of time-shifted redundant circuits and replica circuits. Measurement-based experiments using a fabricated test chip are performed. For an approximately similar false positive error probability for time-shifted redundant circuits and replica circuits, the false negative error probability of time-shifted redundant circuits is approximately two orders of magnitude less than that of the replica circuits. When attaining a false negative error of zero, time-shifted redundant circuits achieves one order of magnitude less in false positive error probability than that of the replica circuits.

Keywords-Dynamic voltage variations; voltage scaling; replica circuits; time diversity; timing errors; error prediction; error detection; variation tolerant circuits

I. INTRODUCTION

With the aggressive process scaling, performance variation due to manufacturing variability has become a serious concern preventing chip designers from pursuing high performance with a high timing yield. This variation is classified as static variation since it is determined in the manufacturing process and is constant after the fabrication. On the other hand, the effect of transistor aging and dynamic variation in supply voltage and temperature is also becoming more prominent. For avoiding timing failures due to both static and dynamic variations, design margins are being built into the operating voltage and frequency settings in order to account for the worst case conditions. However, imposing such guard-bands prohibits the user from exploiting the high performance offered by increasing the operation frequency, or the low power consumption offered by reducing the supply voltage.

Traditionally, replica circuits have been used for performance monitoring. Adaptive control techniques with a critical path replica have been presented in [1]–[3]. Recently, [4] applied Tunable Replica Circuits (TRC) to a microprocessor and the performance improvement and power reduction are demonstrated using a 45nm test chip. The replica circuits in the presented techniques, which mimics the critical path, is fully exercised every cycle. This suggests that these Yukio Mitsuyama School of Systems Engineering, Kochi University of Technology & JST, CREST Kochi, Japan Email: mitsuyama.yukio@kochi-tech.ac.jp

techniques aim to sustain a necessary and sufficient timing margin at all times in all the paths. On the other hand, the critical paths are not activated often in most circuits. This means that in most cycles, the timing margins of the activated paths are larger than necessary.

Such a stochastic property on critical path activation is exploited for aggressive power reduction in [4]–[7]. [7] predicts timing error occurrences using timing error predictive flip-flops, and proactive performance adaptation is executed. On the other hand, [4]–[6] detect timing errors using Error Detection Sequentials (EDS), which are also known as Razor flip-flops. Upon detecting a timing error in the critical path, re-execution and recovery occur. EDS is a flip-flop accompanied by an extra latch that also captures the data slightly after the system clock. The outputs of the main flip-flop and the latch are compared to check whether a timing error occurs or not. As the output data arrives later than the system clock edge, the outputs become different, and the EDS detects a timing error.

Recently, timing error prediction using TRC is utilized not only for coping with static manufacturing variability but also for overcoming dynamic variations [4]. For example, the TRC circuit could be designed so that it fails when a dynamic voltage droop occurs inducing a timing error. The TRC consisting of different logic gates such as inverters, NAND, and other elements needs to mimic the critical paths. Although the longest critical path varies depending on manufacturing variability, the delay of TRC must be larger than that of the longest path in all the variational device parameter space. Furthermore, for handling dynamic voltage variations, the sensitivity of the TRC to supply voltage should be the same as that of the real circuit. Due to such a difficulty, it is used only for high-end processors. Heavy tuning is required to calibrate the sensitivity of the TRC to track critical path delays of the real circuit.

On the other hand, when it comes to reconfigurable devices, implementing replica circuits has a higher compatibility since the behavior of critical path of the real circuit can be exactly mimicked by simply mapping the critical path of the circuit again on the available resources, and heavy tuning of the replica circuit is no longer necessary since the exact circuit is replicated. In addition, in case of coarse-grained reconfigurable device, the number of critical path candidates is quite limited, and just a simple tuning to pad a small delay is sufficient. Thus, the mismatch problem between the replica and the real circuit is less significant. On the other hand, as mentioned above, the critical paths are not often activated, and hence most of the timing error predictions by the replica circuit are false positive. This false positive prediction could significantly degrade the application throughput due to the frequent recovery through re-execution.

An approach that reduces false positive prediction while keeping the false negative low is utilizing time redundancy. The time redundancy is one of the concurrent error detection techniques, and it is often adopted in reliability-demanding applications [8]. The same computation is executed after a certain period of time, and the results are compared. The advantage is that the stochastic property on critical path activation mentioned above can be exploited and the false positive prediction can be significantly reduced when compared to the replica circuit. Time-shifted redundant circuits are also compatible with reconfigurable devices where they can be implemented easily.

Nevertheless, in the field of reconfigurable devices such as Field Programmable Gate Arrays (FPGAs), sub-optimal approaches such as imposing guard-bands are still utilized to accommodate the Process, power supply Voltage, and Temperature (PVT) variations, and the maximum allowable clock frequencies reported by vendor synthesis, placement and routing tools are significantly lower than the ones that the device can really deliver [9]. Although there is ongoing research on the performance reliability and improvement of reconfigurable devices [10], [11], none demonstrate the value and effectiveness of replica circuit for PVT. This is the first work that introduces replica circuits in reconfigurable devices addressing PVT variations as far as the authors have investigated.

Motivated by the compatibility of time-shifted redundant circuits and replica circuits with reconfigurable devices, this work introduces replica circuits in reconfigurable devices, and evaluates the effectiveness of using time-shifted redundant circuits and replica circuits in detecting and predicting timing error occurrences in a real circuit mapped on a reconfigurable device. An application is implemented on a coarse-grained reconfigurable architecture developed in one of our previous works [12], and measurement-based experiments using a fabricated test chip are performed. Comparison results show how much power reduction and throughput improvement can be obtained by exploiting low probability of critical path activation, and the trade-offs offered by each approach in terms of performance, power dissipation, and error probability are presented. The immunity to dynamic voltage variation is compared, and the utility of time redundancy is presented.

II. REPLICA CIRCUIT AND TIME REDUNDANCY

A. Replica circuits

The critical path replica circuit is designed such that the delay of the replica circuit is larger than that of the longest path in the real circuit. It is noteworthy that this relation should hold even under static manufacturing variability and dynamic voltage variation, taking into account the fact that the longest path varies according to these variations. A performance adaptation technique such as supply voltage scaling is then applied, and the supply voltage is reduced to the voltage at which the replica circuit is at the edge of causing timing errors. Here, if the delay of the replica circuit is much larger than the longest path delay in the real circuit, the delay of the real circuit is overestimated resulting in unnecessary performance loss. On the other hand, if the delay of replica circuit is smaller than the longest path delay in a certain operating condition, timing errors in the real circuit will occur. Thus, depending on the replica circuit design, the achievable circuit performance and timing failure probability vary.

This replica circuit can be also used for dynamic variation tolerance by exercising the replica circuit every cycle [4]. When the replica circuit causes a timing error, there is a possibility that timing errors happened in the real circuit. Once a timing error happens in the replica circuit, the states of the real circuit are recovered and the computation is reexecuted. It is likely that the operating condition of the second computation is different from that of the first one, and the second computation might succeed. Another option for ensuring the success of the re-execution could be adopting a multi-cycle operation for the problematic cycles. On the other hand, the critical paths in the real circuit are not activated so often, while the replica circuit is excited every cycle. This means that even when timing errors occur in the replica circuit, timing errors do not necessarily occur in the real circuit. When the sensitivity of the replica circuit to the supply voltage droop is different from that of the critical paths in the real circuit, there is a possibility that the replica circuit does not cause a timing error even when the real circuit causes a timing error. On the other hand, if the delay of replica circuit is much larger than that of the real circuit, timing errors in the replica circuit might appear more frequently. The timing errors that occur solely in the replica circuit involve unnecessary recovery and re-execution, and hence they degrade application performance and throughput. Thus, depending on the replica circuit design, the application performance and timing failure probability under dynamic variation also vary.

In discussing the error detection, we need to consider four situations, as listed in Table I. The first situation (true negative) is that where both the replica and the real circuits

Table IFOUR SITUATIONS OF REPLICA CIRCUITS.

Situation	Replica circuit	Real circuit
True negative	Correct	Correct
False positive	Timing Error	Correct
False negative	Correct	Timing error
True positive	Timing error	Timing error

work without timing errors, which corresponds to the normal operation. In the second situation (false positive), the replica circuit produces a timing error, while the real circuit does not. In this case, recovery is performed, decreasing the throughput depending on the number of the cycles necessary for the recovery phase. On the other hand, a false negative situation arises when the replica circuit does not produce a timing error while the real circuit does. The false negative error is not desirable for reliable applications, and a value converging to zero would be most appealing. In order to reduce the false negative errors, the delay of the replica circuit must be increased, failing more frequently. However, it increases the probability of the false positive, and sacrifices the application throughput and circuit speed.

B. Time redundancy

Concurrent error checking is a popular approach for fault tolerance, and time redundancy is one of the concurrent error checking techniques. For dynamic voltage variation tolerance, the concept of time diversity is effective. The same error outbreak can be prevented by diversifying the time of operation of redundant computation. Figure 1 illustrates the time diversity and its effect. The functionalities of functional blocks A and B are identical, the input data is temporally shifted, and the outputs for the same input data are compared in order to check for errors. In this illustration, the two functional blocks are processing different data at the time the electrical noise occurs, and the outputs of these two blocks are affected in different ways. If inconsistency is found in the outputs, we notice that there was an error occurrence, and accordingly we perform recovery and re-execution. An important factor of time redundancy is the amount of shifting time **d**, which characterizes the immunity to the dynamic variation. As **d** becomes larger, temporally-long variations can be tolerated, however, the additional hardware necessary for time shifting increases. In contrast, when d is small, for example just a single clock cycle, the noise spreading in successive two cycles might not be tolerated; i.e. the same timing error might occur in both circuits causing identical incorrect output.

In case of time redundancy, we have three situations, as listed in Table II. When the outputs of the two blocks are identical, there are two situations; true negative and false negative. This false negative occurs when both blocks cause the same timing error. In case of time redundancy, both blocks are equally treated, and hence the inconsistency of the outputs is regarded as the timing error in the real circuit.



Figure 1. Diversify the operation timings of redundant computation.

Table II Three situations of time redundancy.

	Outputs	Real circuit		
True negative	Identical	Correct		
False negative	Identical	Timing error		
False positive	Different	Timing error		

Therefore, false positive is not included in the table. Thus, to reduce the probability of false negative, we need to intensify time diversity and decrease the probability that both the circuits experience the same timing error.

III. RECONFIGURABLE ARCHITECTURE OVERVIEW

Figure 2 illustrates the overview of the reconfigurable architecture used in this study [12]. Clusters, which are basic elements of this architecture, are placed repeatedly in a two-dimensional array. Each cluster contains four cells connected to each other using diagonal tracks or connected to the cells in the adjacent cluster using horizontal and vertical tracks. The architecture also posses some embedded flexible reliability mechanism, but they will not be explained as they are out of the scope of this paper. Each cell can perform addition and subtraction operations with or without cooperation of the neighboring cells in the same cluster. It also can be configured to perform logical operations such as logical AND and OR, multiplexing, and fixed or variable shifting. A 4×8 cluster array was fabricated in a 65 nm CMOS process and the design is depicted in Fig. 3, and its micro-graph is shown in Fig. 4. This prototype chip will be used throughout the study in this work.



Figure 2. Reconfigurable architecture overview.



Figure 3. Reconfigurable architecture prototype chip.



Figure 4. Micro-graph of prototype chip.

IV. EXPERIMENTAL SETUP

The setup of the experiment is shown in Fig. 5. It consists of a PC, a test board containing a Stratix II FPGA connected to the reconfigurable architecture prototype chip. The application chosen to be mapped on the reconfigurable architecture is a 4-tap FIR filter. One of the possible implementations is shown in Fig. 6. The constants of the mapped FIR filter were all identical and equal to 0.25, which was represented by a 2-bit shift to the right. On the prototype chip, an FIR filter, a time-shifted FIR filter (TFIR), and the replica circuit (RC) were mapped. Similarly, on the FPGA an FIR, RC, a shift register, and a comparator circuit were implemented. The circuits on the FPGA are responsible for generating the correct outputs for comparison purposes. Note that both the FPGA and the prototype chips are necessary since timing errors would arise in the prototype chip as it undergoes dynamic and static supply voltage variation, and the FPGA is used for detecting such errors.

In order to implement the replica circuit, the critical path of the FIR filter was identified and is highlighted with the thick line showing the flow of data on it in Fig. 6. It is noteworthy that the delay of the replica circuit should be equal or larger than that of the critical path of the real circuit in order to make sure that replica circuit fails to produce output when the real circuit does. Knowing this, three different replica circuits were prepared with different delays added to the critical path length of the real application : 148 ps, 308 ps, and 468 ps. The delays were generated



Figure 5. Setup of the experiment.



Figure 6. 4-tap FIR filter.

by using the routing resources on the chip. They were estimated from design data, not from measurement results. When the delay is larger than 468 ps, the replica always fails at nominal voltage without any dynamic noise. As for the time-shifted FIR filter, the key implementation parameter is the amount of cycles by which the FIR is shifted, and it characterizes the diversity of dynamic operating condition, as mentioned in Section II-B. In this study, 1-, 2-, 4- and 8-cycle time shifting are considered.

The PC generates random inputs for the FIR circuits. We also need to give an input pattern to RC. This input to the RC has to be carefully selected in order to exercise the critical path every cycle. The comparator circuit compares the outputs of the chip with the generated outputs of the RCs and the FIRs. It should be noted that the main purpose of the RC on the FPGA is to just produce the expected output in order to compare with the output of RC circuit on the prototype chip.

In this study, two types of experiments are performed and are as follows

A. Static variation experiments

In the static variation experiments, the supply voltage is set to different values between 1.0 V to 1.2 V, and the output of the RC and FIR are monitored and compared. This range of supply voltage includes voltage over-scaling accompanied with timing-violating paths. For each supply voltage value,



Figure 7. Noise injection circuit and pulse waveform.

the comparisons are performed for 2 minutes, allowing approximately 1.6×10^9 comparisons at a 26 MHz operating frequency. In these experiments, there is no significance in changing the delay of the time-shifted FIR, therefore, it will be fixed to 1 cycle delay, and it will be referred to as FIR replica circuit (FIR RC) in static variation experiments. This FIR RC is very important as it serves as a replica for the FIR with zero delay on the critical path, and it also takes advantage of the stochastic probability of the critical path activation, unlike the RC circuits.

B. Dynamic variation experiments

In the dynamic variation experiments, dynamic noises are periodically injected to a 1.2V supply voltage through a capacitor by a pulse generator, as shown in Fig. 7(a). The waveform at the output of the pulse generator is shown in Fig. 7(b). The dynamic noise appears at the rising and falling edges. The amplitude at the pins of the chip package can be manipulated by varying V in Fig. 7(b). V was set to 1V, 2V, and 3V causing a variation of Δ 130mV, Δ 210mV, and Δ 270mV, respectively at the pins of the chip package. Approximately 16×10^9 comparisons were executed in a period of 20 minutes each. In dynamic variation experiments, timeshifted FIR will be referred to as TFIR-1, TFIR-2, TFIR-4, and TFIR-8.

V. RESULTS AND DISCUSSION

A. Static variation

The results of static variation experiment for the replica circuits are shown in Fig. 8 and Fig. 9. The notation of RC 148 ps denotes the results of the RC with 148 ps additional delay. The ideal error detection RC is a virtual circuit that ideally produces timing errors only when there are timing errors in the real circuit, which means false negative and



Figure 8. False positive probability obtained from static variation experiment of RC.



Figure 9. False negative probability obtained from static variation experiments of RC.

false positive are zero. It is demonstrated as the reference of the upper bound performance.

Figures 8 and 9 demonstrate the importance of selecting the appropriate replica circuit for the designated application. Choosing an RC circuit with larger delay will increase the number of false positive errors while minimizing the number of false negative ones in the real circuit. It involves the increase in the number of recovery cycles, and thereby reduces the throughput. Looking at Fig. 8, we see that the FIR RC and the RC 148 ps offer approximately the same false positive probability. However, by looking at Fig. 9, we see that the false negative error probability of RC 148 ps is 3.336×10^{-3} and that of FIR RC is 1×10^{-5} at 1V. That is approximately two orders of magnitude more when compared to FIR RC. This means that FIR RC reduces false positive errors while keeping the false negative error count low. Remind that the main advantage of FIR RC from RC is the active use of stochastic property of the critical path activation, and this two orders of magnitude difference comes from the utilization of this property.



Figure 10. Example demonstrating the effect of false negative probability on throughput.



Figure 11. Example demonstrating throughput vs. normalized dissipated power for each circuit.

For understanding the significance of the false negative probability shown in Fig. 9, we study the effect of this probability on the overall throughput of the circuit. In Fig. 10, we assuming that the number of recovery cycles for each RC failure is 10, 1k, and 100k cycles. We show the overall throughput versus the supply voltage for the FIR RC. For a supply voltage of 1.05V, the throughputs for recovery phases of 10 cycles, 1k cycles, and 100k cycles are equal to 1.59×10^7 , 3.63×10^5 , and 3.68×10^3 which have a difference of approximately two orders of magnitude from one to the next.

We next demonstrate the trade-off between the throughput and power. Assuming that the number of recovery cycles for each predicted error is 100, and that the re-execution will be correct, then throughput versus the normalized dissipated power is shown in Fig. 11. The normalized power dissipation is calculated as a first order analysis assuming it is proportional to supply voltage squared.

It can be observed that there is a considerable tradeoff between the normalized power and throughput for each circuit. In addition, the curves are different depending on the error detection circuits, as we expected. If a throughput of 4.9×10^6 is required, then the minimum normalized dissipated power for the RC 148 ps, FIR RC, RC 308 ps, and RC 468 ps is 1.1, 1.3, 1.16, and 1.26, respectively.

By constraining the false negative probability to less than 0.0005 as an example, the throughput versus normalized



Figure 12. Example demonstrating maximum throughput vs. normalized dissipated power of possible solutions that satisfy a constraint of false negative error probability ≤ 0.0005 .

dissipated power of the possible solutions that satisfy the constraints is shown in Fig. 12. This figure shows that between a range of normalized powers of 1.04 - 1.44, the FIR RC and the RC 148 ps can be used. Below 1.04, the error probability for the RC 148 ps becomes larger than 0.0005, thereby switching to the RC with the next larger delay (RC 308 ps).

B. Dynamic variation

In the experiments of dynamic variation, setting the V to 1 V and 2 V did not induce any timing errors. However, results for setting V to 3 V causing a variation of $\Delta 270$ mV in the supply voltage produced several significant results. The $\Delta 270$ mV noise signal taken by the oscilloscope is shown in Fig. 13. The results are summarized in Table III. The RC 468 ps and the TFIR circuits offer 0 false negative errors, however, the TFIR circuits offer a probability of false positive that is less by approximately one order of magnitude than that of the RC 468 ps.

Figure 14 shows the throughput and the false negative count for each scheme. The RC 148 ps shows the highest false negative error count. Focusing on the near zero false negative error count, we can see that the TFIR circuits offer a higher throughput when compared to the RC 468 ps.

In Fig. 14, the TFIR improves the throughput up to 4cycle shifting (TFIR-4), which can be explained by the time diversity as explained in Section II-B. On the other hand, in case of 8-cycle shifting (TFIR-8), the throughput becomes comparable to that of TFIR-2. In this experiment, we solely observed the noise waveform outside the chip as shown in Fig. 13, and the noise that is actually given to the FIR filter inside the chip cannot be assessed. This non-monotonicity in throughput might be related to the resonance frequency on the chip, and 8-cycle shifting is affected by the resonance. Further study is necessary, but TFIR-4 should be selected for tolerating dynamic voltage variation in this test case.

 Table III

 FALSE NEGATIVE COUNT AND FALSE POSITIVE PROBABILITY.

	RC 148 ps	RC 308 ps	RC 468 ps	TFIR-1	TFIR-2	TFIR-4	TFIR-8
False negative error count	6962	4668	0	0	0	0	0
Probability of false positives	0.00217	0.0094	0.0133	0.00332	0.00316	0.00033	0.0039



Figure 13. Noise signal induced from a pulse of 3 V in amplitude.



Figure 14. Throughput and false negative count collected in the dynamic noise experiments.

VI. CONCLUSION

A comparative study on performance and timing failure probability between time-shifted redundant circuits and replica circuits has been demonstrated. Measurement-based experiments using a fabricated test chip are performed. It is shown that the time-shifted redundant circuit reduces false positive errors while keeping the false negative error count low. When having an approximately similar false positive error probability, false negative error probability of time-shifted redundancy is approximately two orders of magnitude less than that of the replica circuits. For a false negative error of zero, time-shifted redundancy achieves one order of magnitude less in false positive probability than that of the replica circuits.

VII. ACKNOWLEDGMENT

The VLSI chip in this study has been fabricated in the chip fabrication program of VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with eShuttle Inc. and Fujitsu Semiconductor Ltd. This study is partially supported by NEDO.

REFERENCES

- [1] T. Kuroda, K. Suzuki, S. Mita, T. Fujita, F. Yamane, F. Sano, A. Chiba, Y. Watanabe, K. Matsuda, T. Maeda, T. Sakurai, and T. Furuyama, "Variable supply-voltage scheme for low-power high-speed CMOS digital design," IEEE J. Solid-State Circuits, vol. 33, no. 3, pp. 454–462, Mar. 1998.
- [2] J. W. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," IEEE J. Solid-State Circuits, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.
- [3] J. T. Kao, M. Miyazaki, and A. P. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," IEEE J. Solid-State Circuits, vol. 37, no. 11, pp. 1545–1554, Nov. 2002.
- [4] J. Tschanz, K. Bowman, S. Walstra, M. Agostinelli, T. Karnik, and V. De, "Tunable replica circuits and adaptive voltagefrequency techniques for dynamic voltage, temperature, and aging variation tolerance,", Symposium on VLSI Circuits Digest of Technical Papers, 2009.
- [5] S. Das, D. Roberts, S. Lee, S. Pant, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "A self-tuning DVS processor using delay-error detection and correction," IEEE J. Solid-State Circuits, vol. 41, no. 4, pp. 792–804, Apr. 2006.
- [6] S. Das, C. Tokunaga, S. Pant, W. H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. Blaauw, "Razor II: In situ error detection and correction for PVT and SER tolerance," IEEE J. Solid-State Circuits, vol. 44, no. 1, pp. 32–48, Jan. 2009
- [7] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Adaptive performance compensation with in-situ timing error predictive sensors for subthreshold circuits," IEEE Transactions on VLSI Systems, vol. 20, no. 2, pp. 333–343, Feb. 2012.
- [8] N. Kanekawa, E. Ibe, T. Suga, and Y. Uematsu, "Dependability in electronic systems," Springer, 2011.
- [9] J. Freijedo, L. Costas, J. Semiao, J.J. Rodriguez-Andina, M.J. Moure, F. Vargas, I.C. Teixeira and J.P. Teixeira, "Impact of power supply voltage variations on FPGA-based digital systems performance," Journal of Low Power Electron., vol. 6, pp. 339–349, Aug. 2010.
- [10] A. Kumar and M. Anis, "FPGA design for timing yield under process variations," IEEE Transactions on VLSI Systems, vol. 18, no. 3, pp. 423–435, Mar. 2010.
- [11] H. Yu, Q. Xu and Ph. H.W. Leong, "Fine-grained characterization of process variation in FPGAs," in Proc. International Conference on Field Programmable Technology, pp. 138–145, Dec. 2010.
- [12] D. Alnajjar, Y. Ko, T. Imagawa, H. Konoura, M. Hiromoto, Y. Mitsuyama, M. Hashimoto, H. Ochi, and T. Onoye, "Coarsegrained dynamically reconfigurable architecture with flexible reliability," in Proc. Field Programmable Logic and Applications Conference, Sept. 2009.