Delay-Fault Tolerance to Power Supply Voltage Disturbances Analysis in Nanometer Technologies

J. Semião¹, J. Freijedo², J. Rodriguez-Andina², F. Vargas³, M. Santos⁴, I. Teixeira⁴, P. Teixeira⁴

¹ Univ. of Algarve, Instituto Superior de Engenharia, Campus da Penha, 8005-139 Faro, Portugal, jsemiao@ualg.pt
² Univ. of Vigo, Dept. de Tecnología Electrónica, Campus Universitario, 36310 Vigo, Spain, {jfreijedo, jjrdguez} @uvigo.es
³ PUCRS, Electrical Engineering Dept., Av. Ipiranga, 6681, 90619-900 Porto Alegre, Brazil, vargas@computer.org
⁴ IST, Technical Univ. of Lisbon / INESC-ID Lisboa, R. Alves Redol, 9, 3º, 1000-029 Lisboa, Portugal, {marcelino.santos, isabel.teixeira, paulo.teixeira} @ist.utl.pt

Abstract- In nanometer technologies, as variability is becoming one of the leading causes for chip failures, signal integrity is a key issue for high-performance digital System-on-Chip (SoC) products. In this paper, analysis is focused on the occurrence of Delay-faults due to Power-supply disturbances in nanometer technologies. Using a previously proposed VT (power supply Voltage and Temperature)-aware time management methodology, it is shown that nanometer technologies impose the need of fault-tolerance methodologies, although the margins of tolerance or fault-free operations are being reduced as technology scales down. SPICE simulation results with 350nm, 130nm, 90nm, 65nm, 45nm and 32nm CMOS technologies show an increasing dependence of propagation delays on power supply variations, as technology is being scaled down. Monte Carlo simulations show that, even in the presence of process variations, a dynamic delay-fault tolerance methodology can be rewarding even at nanometer scale, although the margins for Power-supply variations are becoming smaller.

I. INTRODUCTION

Technology scaling down to the nanometer range leads to increased process variations and reduced noise margins, making signal integrity a key issue for high-performance digital System-on-Chip (SoC) products [1]. There are many causes for signal integrity loss and chip failures (crosstalk [2], power supply voltage overshoot/undershoot [3], reflection, electromagnetic interference (EMI) [4], power-supply noise [5], and signal skew [6]), but as technology scales down, variability is rapidly becoming one of the leading causes for chip failures and delayed schedules, defined as the unpredictability, inconsistency, unevenness, and changeability associated with a given nuance [7].

Often, in production test, signal integrity and variability problems can influence the test result, especially when delay faults are under scrutiny. In fact, delay faults in synchronous digital SoC may be caused by environmental disturbances or circuit operation. Signals may arrive too late (or too early) at key nodes in the circuit’s structure. Hence, data or control signals are captured and transmitted in the incorrect time window, as the assumed time slack is violated. Such delay faults are not due to a physical defect and are not permanent; typically, they are operation-dependent. Test power may exceed the power consumption in normal mode of operation, which may aggravate the power noise and temperature rise. Clearly, if the product is made more tolerant to such time variations, more chips could pass the test and yield loss could be recuperated. The impact of power supply noise in delay testing is a matter of current research [8].

In the design phase, the time slack is defined taking into account process and power supply voltage variations, which are unavoidable in manufacturing and circuit operation. Worst case conditions take into account both effects, which may significantly degrade circuit performance. For instance, for the target CMOS technology we use, design examples are made to meet time specs with VDD down to 90% of its nominal value. If, at the same time, process variations are considered (as it will be shown in section IV), the maximum clock frequency is significantly degraded. Circuit operation is also responsible for temperature (T) variations, namely causing operation-dependent hot spots. High-performance products require limited time slack. Hence, making by design the circuit more tolerant to VDD and/or T variations, regardless of their origin, will help restricting yield loss.

This paper is focused on delay and power variability in nanometer CMOS circuits, which is influenced by many contributors [9]. The purpose of the work is to analyze the influence of power supply voltage disturbances in nanometer technologies, as a cause for delay fault occurrence, and to analyze a power supply voltage disturbance tolerant methodology when applied to different nanometer technologies, namely, 130nm, 90nm, 65nm, 45nm and 32nm. The methodology aims to enhance circuit tolerance to power supply and temperature disturbances. The local insertion of dynamic delays in specific memory cells, according to VDD and T unwanted disturbances, eases correct data capture and reduces delay-fault occurrence. Local VDD (and/or T) variations are to be expected, due to power grid partitioning and circuit operation.

This paper is organized as follows. In Section 2, a review of the previous work is provided. Section 3 presents the analysis of power supply reduction influence on path delay degradation for several nanometer technologies. Also in section 3 a time borrowing methodology for delay fault tolerance is analyzed, when implemented in nanometer technologies. In Section 4, experimental results based on Monte Carlo SPICE simulations for a pipeline circuit example
II. PREVIOUS WORK

Process variations depend on the semiconductor technology, and on the manufacturing process. This also applies to physical defects, induced during IC manufacturing. However, VT (power-supply Voltage and Temperature) variations are caused either by environmental disturbances (e.g., EMI/EMC [11]), or by circuit operation (switching activity). They induce timing variations not only in the “functional part” of the system (combinational and sequential logic, memory, etc.), but also in the supporting infrastructure – the power grid, and the clock distribution network (CDN). An operation-dependent power and thermal map can be defined, with variable (in time, and space) VDD and T values, modulating power, clock, control and data signal propagation. This is difficult to predict, at the design stage. The behavior of the power grid and of the CDN is very sensitive to power and thermal variations [12].

Therefore, signal integrity problems due to timing variations occur [13]. Variable clock skew, due to operation-dependent VT variations, may have to be taken into account [12]. CDN are designed to act as the fast lane of a highway, while signal paths in combinational logic behave as the slow lanes in such highway. When VT variations occur, a loss of synchronism may occur, associated with data captured too early, or too late, as compared to the correct timing [14].

Typically, how do designers deal with PVT variations? A worst case scenario is considered, where pre-specified cumulative variations are assumed, e.g., $V_{DD} \in (0.9; 1)\times V_{DD_{nom}}$, $T \in (T_o, T_1)$ (commercial, or military specifications) and process variation data of the target IC technology. The maximum frequency of operation ($f_{max}$) is defined, taking into account the time required to propagate signals along the longest critical paths, and adding a time slack. However, worst-case PVT variations scenario leads to unacceptable performance loss, and is unduly pessimistic [9]. Designers need to figure out methods to restrict time slack, thus boosting performance, or search for new design approaches that considers multiple concurrent corner analysis [7]. The Razor technique [10] implicitly looks for minimum time slacks, with the objective of applying DVFS. As we move to new node technologies, performance gains are limited. Hence, a common technique is to partition critical paths propagation delays, by pipelining data processing along combinational modules. Latency is introduced, but performance is enhanced [15].

From a test point of view, test power generally exceeds the power consumption in normal mode, as enhanced switching activity is usually required for structural testing. As a consequence, larger power supply noise is to be expected (due to IR drops), inducing larger VT variations [8]. The variation of metal resistance with temperature (Cu resistance approaching $\Omega$ resistance) leads to enhanced interconnection delays, which may also compromise test results, as at-speed testing is required to detect delay faults. Resistive open and bridging defects need to be detected in production, so 0 DPM (Defective parts per Million) can be reached. Again, delay testing is essential. On-chip availability of variable VDD may facilitate production and lifetime testing. Parametric testing is a good resource for screening out defective parts, and avoiding yield loss due to false positives.

In order to enhance performance, design techniques, such as skew-tolerant domino design and timing analysis algorithm (including different amount of clock skew between different elements) have been introduced [16][17]. The time borrowing (TB) concept has been used in [16][18] to make skew-tolerant circuits operate at higher speed. Basically, a predefined additional time $\Delta t$ is given for signal propagation to a certain segment of the signal path, taking into account that this borrowed time interval can be recovered along the full signal path, as a subsequent segment executes its functionality in less time. This allows the reduction of the time slack, enhancing speed operation.

All these techniques consider static time definition, in the sense that clock signals are pre-defined, and the amounts of borrowed time also pre-computed. For instance, in Razor, the delayed clock signal, applied to the extra latch in each FF, has a static, predefined clock phase. None of these techniques take VT variations into account. Incidentally, structural delay testing of such high-performance circuits is a difficult problem [19]. However, VT variations do occur.

In order to guarantee correct timing performance in the presence of unwanted VT variations, the ultimate solution would be to increase the time slack (worst case scenario), i.e., to reduce the clock signal frequency. However, performance degradation cannot be tolerated. Performance specifications must be complied with. Nevertheless, VT variations are not permanent, they are operation-dependent and, typically, their ability to induce errors is scarce in time and place. What would be required is to have the possibility of on-line monitoring such variations, where and when they occur, and dynamically introducing extra time to signal capture ($\Delta t (V_{DD}, T, x, y)$). This needs to be performed with low overhead. Time recovering, along the full signal paths must also be guaranteed. The goal is clever time management, by performing dynamic time borrowing and recovery.

Recently, the authors proposed a first methodology [13][20] that uses clock duty cycle modulation (CDC) to adjust the instant of data capture, in the presence of VT variations. In [20] a CSL (Clock Stretching Logic) block is used to modulate CDC and to enhance circuit tolerance without performance degradation. However, many practical synchronous circuits have simultaneous rising and falling edge triggered memory elements. This makes mandatory the use of complementary CSL blocks, which limits its usefulness.

To overcome this problems and still enhance circuit tolerance without degrade performance, the authors recently proposed a methodology ([21][22]) to control and adjust the clock phase (instead of the clock duty-cycle) of pre-determined memory cells’ clock signal (the critical memory cells), and thus enhancing tolerance in the presence of VDD and T variations. The work presented in [21] is focused on the design of a mixed-signal cell, the Dynamic Delay Buffer...
(DDB) cell, that senses VDD and T variations and changes clock phase accordingly. In reference [22], the work was updated focusing now the application of the methodology to pipeline circuits and explaining how the improvements in tolerance can be obtained. There is also the application of the methodology to some circuit examples and SPICE simulation. In [23], the main new contributions were: (1) the analysis of the methodology behavior with process variation; (2) the resolution of two methodology limitations previously identified (and now implemented and analyzed with process variations); (3) preliminary results on a proprietary software tool (Dynamic Delay Analyzer (DyDA)), to automatically identify the critical memory cells and calculate the DDB insertion points; and (4) the analysis of delay fault detection in a circuit with enhanced tolerance to VDD and T variations.

In authors’ previous work, the delay fault tolerant methodology is widely explained and experimented. However, the AMS CMOS 0.35 μm (C35B4) digital technology and design kit were used. All SPICE simulations and the two test chips developed use the referred 350nm old technology. But how does a time borrowing methodology work in a nanometer technology, when variability is widely increased and noise margins are reduced? What is the expected yield loss reduction in new technologies for the delay-fault tolerant approach? The present work aims to answer these questions, and the new main contribution is the analysis of authors’ previous proposed methodology implementation to nanometer technologies, namely to 130nm, 90nm, 65nm, 45nm and 32nm technology.

III. DELAY-FAULT TOLERANCE IN NANOMETER TECHNOLOGIES

The methodology presented in authors’ previous work [22][23] use a dynamic time borrowing strategy to dynamically delay the clock signal of specific memory cells, providing additional time to accommodate the increased signal propagation delay through the logic whose VDD or temperature is being disturbed. This allows controlling the instant of the active clock edge trigger for the memory cells with the smallest time slack margin. The innovative idea in the methodology is to adapt the clock delay (skew) locally and dynamically, as needed. This is done with the selective insertion of DDB (dynamic delay buffer) cells in predetermined memory cells (critical memory cells) that are sensitive to VT variations and can act as a sensor and as an actuator, introducing the necessary delay (skew) in the clock signal and borrowing time from the subsequent pipeline stage to accommodate the delayed signal, while maintaining at-speed clock rate. More details on the delay-fault tolerant Methodology can be found in [21][22][23].

Time borrowing techniques have been widely used in pipeline based circuits, because of their particular architecture [16][18]. Hence, the methodology presented in this paper is easily applicable to pipeline based circuits. However, as it can be seen on [23], it is possible to achieve improvements in time slack margins on FSM (Finite State Machines) circuits, depending on their structure.

A. Path Delay Analysis with Time Borrowing

To evaluate the behavior of a combinational path to VDD, SPICE simulations were performed using a combinational Critical Path (CP) example (chain of 22 inverters, here referred as “case study 1”). The circuit was implemented with an AMS CMOS 350nm technology. Consider that the referred CP is the combinational CP of a pipeline circuit (see figure 1) working at 667MHz (clock period of 1.5ns). What is the expected improvement (when applying a DDB cell) to dynamically provide additional time to capture the correct data from the CP? Assuming negligible setup and hold times for memory cells (Q0 and Q1 in figure 1), the CP delay will be allowed to increase, further than the clock period, to an upper limit depending on the next clock cycle path delay (ideally, the upper limit is twice the clock period). Fig. 2 shows simulations on the propagation delays in a DDB cell (implemented with 2 additional inverters to restore fast transitions) placed in the clock path of the memory cell that ends the combinational critical path (Q1), and compare them to the CP delay degradation. The vertical lines area on the left indicates when the CP propagation delay is smaller than the clock period. In this case, there are no timing violations. As VDD decreases, timing violations may occur when the CP is greater than the clock period (1.5ns). The initial time slack is \( t_s = 0.18 \text{ns} \) and the DDB intrinsic delay is \( t_{DDB} = 0.338 \text{ns} \). However, using a delayed clock provided by the DDB cell (in critical memory Q1), the timing violations will be eliminated in the VDD \( \in [2.9; 2.3] \text{V} \) range (region represented by the sloped lines area, in the middle), allowing VDD to decrease further than before and thus improving tolerance to VDD variations. The new limit (VDD=2.3V) occurs when CP delay exceeds the modified DDB clock period limit (or when the next clock

![Figure 1: An inverters’ chain creating a combinational critical path (case study 1), here placed in a hypothetical pipeline stage.](image)

![Figure 2: Methodology Improvement with VDD Drop (22 inv. CP, CMOS AMS 350 nm).](image)
cycle starts to have timing violations). For additional $V_{DD}$ drops, performance will collapse (horizontal lines area in the right, $V_{DD} < 2.3V$).

However this was performed with an old technology of 350nm. Using the Berkeley Predictive Technology Models (PTM) [24] for CMOS transistors, the same circuit was implemented in the following nanometer technologies: 130nm, 90nm, 65nm, 45nm and 32nm. The period of operation was defined in the same way as before, that is, the circuit must operate at a depleted $V_{DD}$ of 90% of its nominal value. Figure 3 shows the frequencies of operation for all technologies considered and, as we can see, the frequency rises as technology scales down.

In order to compare the results with the previous simulation, all the delay values were normalized with the corresponding clock period, and the $V_{DD}$ depleted values were normalized with the corresponding nominal value. Spice simulations were performed in all the technologies considered and the result is shown in Figure 4.

From the analysis of Figure 4, it can be seen that CP delay is more sensitive to $V_{DD}$ voltage drops as technology is reduced, as expected. Consequently, the initial time slack is incremented (to have the same initial tolerance to $V_{DD}$ drops) and the improvement margins (tolerance to $V_{DD}$ voltage drops) are reduce as technology scales down. The high variability in small gate length technologies is caused by the increased dependence of path delay from all variables (in this case, the power supply voltage).

In figure 5 it is shown the maximum $V_{DD}$ reduction (with respect to the nominal value of $V_{DD}$) and the minimum initial time slack (with respect to the clock period). Note that the maximum $V_{DD}$ reduction is decreasing, while the initial time slack that accommodate the same initial $V_{DD}$ drop tolerance is increasing. As a consequence, the margins for a delay fault tolerant methodology application in nanometer technology circuits are being reduced. The ultimate limit for delay-fault tolerant methodology application will be when the initial time slack accommodates all the $V_{DD}$ drop margins, leaving no $V_{DD}$ drop margin beyond 10% of nominal $V_{DD}$ drop value. But, is it rewarding in these technologies to consider the time borrowing strategy? In the next section, Monte Carlo simulations will answer this question.

IV. MONTE CARLO SIMULATIONS RESULTS

As variability is rapidly becoming one of the leading causes for chip failures and delayed schedules, Monte Carlo simulations are needed to calculate the expected Yield loss reduction for each $V_{DD}$ reduction value, and analyze the applicability of the dynamic delay-fault tolerant methodology in nanometer technology range.

In the following circuits, it was used the AMS C35b4 CMOS technology, for the 350nm technology implementations, and the Berkeley PTM models to simulate the circuits implemented in the nanometer technologies 130nm, 90nm, 65nm, 45nm, and 32nm.

Consider now a pipeline example with 3 stages, where a combinational critical path (implemented with 23 inverters in chain) located between two stages is followed in the next pipeline stage by a non-critical combinational path (implemented with 9 inverters’ chain), as illustrated in figure 6, and referred as “case study 2”. Considering the operation with $V_{DD}$ down to 90% of its nominal value, the maximum clock

![Figure 3: Maximum frequency of operation for “case study 1”](image1)

![Figure 4: Critical path delay degradation with VDD reduction.](image2)

![Figure 6: “Case study 2” pipeline example circuit.](image3)
Frequency for the 350nm technology implementation is fclk=0.455 GHz. However, this maximum frequency is reduced to fclk=0.4 GHz (12% degradation), if we want all the samples to pass the performance test at a nominal V_{DD} and allow worst-case conditions for process variations. This value is again reduced if we want to enable all the samples to pass the test with a V_{DD} of 90% of its nominal value (fclk=0.357 GHz, an overall 21.5% degradation). Additional performance degradation results from assuming correct behavior within a temperature range (commercial or military specs). In fact, the operation with worst-case conditions reduces performance to accommodate the uncertainty of all variables, taking a share for temperature, another for V_{DD} and another for process variations.

In this example, attention is focused on process variations. Hence, the frequency was set to fclk=0.4GHz. Monte Carlo simulations (100 iterations for each V_{DD} value, with process and mismatch analysis variation) were performed under the Cadence environment to determine how the performance is being degraded when V_{DD} is reduced. To measure this degradation, a Performance Yield (PY) metric was introduced. In this context, it is defined here as the quotient between the number of samples which pass the performance test and the total number of samples:

\[
PY = \frac{n_{\text{good samples}}}{n_{\text{total samples}}}
\]  

In Figure 7 this Performance Yield degradation with V_{DD} reduction is shown for the 3-stage pipeline example. The line with squares represents the normal design, where no (deliberate) skews were introduced in pipeline stages. Using the DDB basic cell (circuit with Dynamic Skews) in the critical path stage flip-flop (as illustrated in figure 6), the performance yield was increased for all different V_{DD}, as shown in the line with triangles. Basically, the introduction of the DDB cell allows us to accommodate both V_{DD} and process variations with the time penalty associated with only V_{DD} variations (normal design).

It is interesting to notice that, if two common balanced inverters were used in the DDB (instead of two unbalanced inverters), the result would be worse, with less improvement in performance yield. This can be referred as “static skews” and is represented in Figure 7 by the line with circles. The purpose of unbalancing the inverters in the DDB cell is to make the sensors more sensitive to VT variations, as it can be seen in [21]. Hence, dynamic skews are less intrusive to the circuit and can be more effective, when compared to static skews.

Nevertheless, unbalancing the inverters in the DDB cell (with dynamic skews) causes also the weakness of its driving capability. To overcome this problem, restore fast transitions at DDB output and/or add an extra clock skew, the DDB basic cell can be buffered, creating static and dynamic clock skews. The results show improvements in the performance yield, represented by the line with diamonds in Figure 7. In particular, for a reduction of 21% in the supply voltage, 29% yield gain is achieved when comparing the proposed methodology with the static skew, and this performance yield increases to 48% when compared to the normal design (with no deliberated skews introduced). The results can also be analyzed in the performance point of view. With skews and at the experimented frequency of operation, it is guaranteed that all the samples pass the test with a V_{DD} reduction of 15%. When compared to the normal design, as mentioned before, only at a reduced clock frequency of f_{clk}=0.357 GHz we could have PY=100% (performance yield) with a V_{DD} reduction of 10%.

Now consider the implementation of the circuit in figure 6 in the following nanometer technologies: 130nm, 90nm, 65nm, 45nm, and 32nm. As it was mentioned, Berkeley PTM CMOS models for these technologies were used to simulate the circuit in the considered technologies. In order to compare the results, only the original circuit (without DDBs in the clock path) and the circuit with dynamic and static skews (the DDB basic cell is followed by a buffer to restore fast transitions) were implemented and simulated.

In order to perform statistical Monte Carlo Simulations with the PTM Berkeley models, independent Gaussian distributions for TOX and L_{eff} parameters were considered. For each technology, Monte Carlo simulations (30 iterations for each V_{DD} value, for 3σ variations with TOX=±4% and L_{eff}=±4%) were performed under the HSPICE (Synopsys) environment to determine how the performance is being degraded when V_{DD} is reduced. The maximum frequency of operation was defined in the same way as before, that is, the circuit must operate with no errors for all the samples in MC simulations. The maximum frequencies obtained for the different technologies are similar to the result obtained in figure 3.

Figure 8 resumes the simulations for the “case study 2” circuit when implemented in the considered technologies. As it can be seen, the tolerance to VDD disturbances is reduced, when reducing the technology. However, the higher sensitivity to VDD variations leads to high yield gains for the same VDD reduction percentage. In particular, except for the 32nm technology, for a reduction of 10% in the supply voltage, 100% yield gain is achieved when comparing the proposed methodology to the normal design (with no deliberated skews introduced).
This result shows us that careful use of time borrowing techniques in nanometer range technologies, may lead to significant yield loss reduction due to delay-faults. However, as circuits are more sensitive to power supply voltage, additional margins should be used when critical memory cells are selected and when time borrowing strategy applicability is analyzed. If small margins of tolerance to VDD variations can lead to high yield gains, small additional errors (unpredicted) can lead to high yield losses.

V. CONCLUSIONS

In this paper the authors analyzed the applicability of a dynamic delay fault tolerant methodology to nanometer range technologies. The methodology, proposed in previous works, aims to increase the delay-fault tolerance of pipeline based digital circuits to power-supply voltage and/or temperature variations in a SoC, using a time borrowing strategy with the insertion of dynamic clock skews in specific memory cells. However, in nanometer technologies, the increasing influence of power supply disturbances in path delay variations makes margins to improve tolerance to tight.

Monte Carlo SPICE simulations show that, for the nanometer technologies considered, namely 130nm, 90nm, 65nm, 45nm and 32nm, it is possible to achieve significant yield gain improvements when the dynamic delay fault methodology is applied to pipeline base circuits. Nevertheless, as circuits are more sensitive to power supply voltage variations in small technologies, the smaller margins to VDD variations impose additional care when methodology application to a circuit is performed.

REFERENCES


