Dynamic Voltage Scaling with Fault-Tolerance for Lifetime Operation

J. Semião
ISE – University of Algarve,
Faro, Portugal
Email: jsemiao@ualg.pt

C. Leong
INESC-ID
Lisbon, Portugal
Email: cl_viper@hotmail.com

R. Cabral, M. B. Santos,
I. C. Teixeira, J. P. Teixeira
IST / INESC-ID
Lisbon, Portugal
Email: \{ruben.cabral, marcelino.santos, isabel.teixeira, paulo.teixeira\}@ist.utl.pt

Abstract—This work consists in a dynamic voltage scaling methodology, with Fault-Tolerance capability, to be used during circuits’ lifetime operation. Methodology implements on-line circuit monitoring, restricting power consumption by adjusting dynamically the power-supply voltage to the minimum value that prevent errors. Error free operation is achieved by using a cooperative work of global and local sensors, sensitive to PVT and Aging (namely, Bias Temperature Instability effects), and by adding an additional fault-tolerance margin to improve error avoidance. Spice simulations in a 65nm CMOS technology demonstrate the results for an example of a dynamic voltage scaling strategy.

Keywords—Dynamic voltage scaling; Fault-Tolerance; Performance sensors; PVTA variations; Bias Temperature Instability.

I. INTRODUCTION

New technologies impose serious challenges on circuits’ reliability and on test methodologies. Process, power-supply Voltage, Temperature and Aging (PVTA) [1], are examples if key parameters that affect transistors’ operation. Variation on these and other parameters, regardless of their origin, may lead to chip failures, especially when several effects occur simultaneously or when cumulative degradations pile up.

Power consumption is also a major topic in today’s circuits, especially in battery operated devices. Interestingly, although power density is increasing (as the areas shrink and the frequencies increase), power supply voltages stop scaling down significantly below 1 volt [5]. Dynamic power is still the main concern, however from 65nm and below, leakage power may become similar [6]. Standard techniques like clock-gating are being used to reduce dynamic power in recent server-class microprocessors [5]. However, aggressive techniques to reduce power consumption are gaining their way, like Dynamic Voltage and Frequency Scaling – DVFS [7]. Moreover, sub-threshold voltage operation is gaining importance in applications where power is critical and performance can be lightened. Sub-threshold digital circuit design has emerged as a low-energy solution for applications with strict energy constraints [9].

This work presents a proof of concept for a DVS scheme, with the purpose to reduce power-supply voltage to the minimum, but still avoiding errors. As circuits age during their life-cycle, so thus the optimal power-supply voltage needs to be constantly monitored. This is done with the cooperative work of global and local performance sensors, designed to be aging aware and focusing the Bias Temperature Instability effect (the negative, NBTI, and the positive, PBTI). These sensors work predictively, i.e., signalizing the need for corrective actions in the VDD when a minimum safety margin defined by design is reached, and thus, preventing errors from happening. Furthermore, apart from the minimum safety margin used, there is also a second time margin, which introduces error-tolerance functionality. This tolerance margin is placed after the clock edge-trigger, providing, if needed, an additional time window to capture a delayed data signal (and borrowing time from subsequent cycles).

II. DYNAMIC VOLTAGE SCALING METHODOLOGY

The Dynamic Voltage Scaling (DVS) block diagram is depicted in Fig.1. It can be identified: (1) a digitally Controlled Power-supply voltage source, to allow voltage scaling in the
circuit under test (CUT) and remaining circuitry; (2) a DVS Controller, to control all DVS operation; (3) a Global Sensor, to perform the on-line monitoring of CUT’s timing response, using dummy critical paths; and (4) the CUT with predetermined flip-flops already replaced for local sensors (with tolerance capability), to allow circuit’s critical paths’ delay monitoring, locally, where errors occur.

The methodology starts with an initial sensor’s calibration procedure in test mode, to tune the global sensor (GS) output with the local sensors’ (LS) error-free operation. This procedure, uses off-line tests (e.g., scan-based delay-fault oriented tests), performed repeatedly for different values in the power-supply voltage, and obtains the minimum \( V_{DD} \) voltage for which no LS is activated (with a controlled error margin defined by design). This information will be stored in a register (\( V_{DD} \) register) and corresponds to the target optimal power-supply voltage. Furthermore, the GS output for this optimal voltage also defines the safe/optimized output for GS, and this information is stored (in GSOSafe registers) and will be used during the on-field operation.

In an on-field operation (normal circuit operation), GS is responsible for the coarse grain delay fault prediction, by comparing the safe/optimized GS output (stored) and the actual GS output. LS will perform a fine grain fault prediction, by triggering a sensor tuning during on-line operation when a LS is activated (and no GS error is detected). Both LS and GS can trigger \( V_{DD} \) voltage corrections. The GS activation period is defined according with the available time slack (the shorter the time slack, the shorter the GS’s activation should be).

Fig.2 depicts the simplified functionality of the DVS controller (and methodology operation). As LS are always on-line, the reset state is “Local Sensing”. “Global Sensing” is triggered when an error is detected by LS, or by a timer. If an error is not detected by LS, neither by GS, no action is taken and it returns to the “Local Sensing”. However, if a global error or a local error is detected, the state is changed to “Change \( V_{DD} \)” state and the \( V_{DD} \) is updated (increased or reduced). If only a global error was detected, the state is changed to “Local Sensing”. Though, if only the LS detected an error, it means that at least one of the CUT’s (circuit under test) CP ages faster than the dummy CP, and so the sensors’ output for safe/optimized operation should also be updated, in the “Update GSOSafe” state.

If there is a possibility that aging degradation is higher in circuit paths than in dummy CPs, the existence of local sensors operating on-line ensures that the system can learn from the sensors’ information and will adapt to the new operating conditions during its lifetime. Moreover, the additional tolerance to delay-faults presented in the local sensors provides an additional safety margin to avoid errors, even if local sensors are not activated during the on-line operation.

### III. LOCAL SENSOR ARCHITECTURE

The proposed LS architecture was previously published in reference [10], as the Scout Flip-flop, a local performance sensor with tolerance capability. The LS is based on a master-slave flip-flop (FF) and intends to constantly observe, monitor or inspect the FF data and inform if an unsafe data transition occurs. The unsafe data transitions are here identified as error-free data captures in the FF that occur in the eminence (within a pre-defined safety-margin) of a delay error. Hence, a virtual window created by design specifies this safety margin to identify the unsafe transitions, consisting in the predictive detection of delay faults. Moreover, a second virtual window is also created to identify the delay-fault tolerance margin of the LS FF. This tolerance is created by delaying data captures in the master latch of the FF, thus avoiding error occurrence in the FF (during the tolerance window) if a late arrival data transition occurs. Therefore, the LS FF includes performance sensor functionality, with additional tolerance and predictive detection of delay faults.

Fig. 3 presents the LS’s architecture and Fig. 4 depicts the two virtual guard band windows created from DE_1 and DE_2.
delays. Note that the word “virtual” is used here because there are no signals with the explicit windows defined. Hence, the delays used allow to tolerate and detect transitions in the data input D of the FF that occur during the specified time frames. An important feature in the LS is that these delays increase with aging, increasing the tolerance and detection margins of the LS as the LS ages itself. This means that the sensitivity of the LS improves with aging, as it is proven in previous publications [2], [3], [8] and [10].

For further information on the LS functionality and architecture, please refer to reference [10]. For further information on the LS insertion criteria and hardware overhead restriction, please refer to references [2], [3] and [8].

IV. GLOBAL SENSOR ARCHITECTURE

Fig. 5 presents the GS architecture, with two dummy CP to produce several delays’ replicas. With the knowledge of the critical paths of the circuit, it is possible to create two fictitious paths, with propagation delays higher than the expected circuit’s CP during its lifetime. One chain is implemented with NOR gates (dummy critical path 1 in Fig. 5), creating a fictitious critical path which will, presumably, age more than the critical paths of the circuit when subjected to the NBTI effect (which strongly influences the degradation of PMOS transistors’ Vth). The other chain is implemented with NAND gates (dummy critical path 2 in Fig. 5), creating another fictitious critical path which will, presumably, age more than the critical paths of the circuit when subjected to the PBTI effect (which strongly influences the degradation of Vth in NMOS transistors).

The NORs and the NANDs input port-map are important for the high aging degradation of the PMOS and NMOS transistors in the respective chain paths. The internal structure of NOR and NAND gates are presented in Fig. 6. For the NOR gate (Fig. 6(a)), the probability for the transistor P1 to be in stress mode is equal to the probability of having its NOR chain input at logic value 0. If global sensor is activated periodically, this signal will most likely be at low logic value most of the time, making a high degradation probability for transistor P1. However, the probability to put P2 in stress mode is equivalent to the probability of having both P1 and P2 transistors on, i.e., P([NOR_\text{chain}]_{\text{input}} = 0) \times P([\text{Age_enable_1}]_{\text{input}} = 0). Considering that global sensor is activated periodically, Age_enable_1 signal has low probability to be at 0 logic value. Henceforth, P2 will have negligible degradation. Yet, a high degradation probability of CP delay’s replica is guaranteed with the high degradation probability of all the P1 transistors of the NOR chain. Moreover, if a higher degradation is needed in the dummy critical path, a 3-input NOR gate can also be used, with 2 of its inputs connected to the same NOR chain input and having now 2 PMOS transistors in a high aging state. Regarding the NAND gates, a similar analysis can be drawn to create a high aging probability in the dummy critical-path 2.

The GS operation is as follows. When the local control unit in the GS (the Global Sensor Controller block) receives a signal to start the analysis of the circuit’s performance and generates control signals (Test_data_1, Age_Enable_1, Test_data_2, Age_Enable_2) to operate the overall performance analysis. These control signals allow to place transparent NOR and NAND gates chain inputs, using signals Age_Enable_1 and Age_Enable_2, so that through signals Test_data_1 and Test_data_2 can be generated a test sequence.
that stimulates the two state transitions at the outputs of the gates of the two chains. Along the two chains, special sensor cells, build with a Latch, a Delay Element and a Stability Checker, are connected at the output of several NAND and NOR gates to create several fictitious paths with different propagation times. The architecture of these sensor cells is defined in [8].

V. SIMULATION RESULTS

For LS results please refer to [10], while for GS results please refer to [8] and [11]. In this paper the purpose is to present complete system simulation results and to make the proof of concept for the DVS methodology. Simulations were performed with HSPICE for the CUT (PM4-2), a 4-bit, 2-stage Pipeline Multiplier with 9 inputs (8 data and 1 clock), 8 data outputs in PTM 65nm [4], with a nominal VDD of 1.1V with 27°C for nominal temperature.

For DVS results, the complete circuit defined in Fig.1 was implemented and a Digitally Controlled Power-supply Voltage Source was designed to allow a dynamic control of circuit’s VDD. It is out of the scope of this work to present the architecture and functionality of the digitally controlled power-supply voltage source. Though, it is a complex mixed-signal CMOS circuit design, with 3 control bits and producing output voltages from 0.7V to 1.1V. SPICE simulations were performed for the complete circuit and the results are summarized in Fig. 7 and Fig. 8. The first three graphs of each figure represent the following 3-bit digital words in decimal values: GAS is the output of the GS; GSO_out is the optimal GS output (tuned during off-line test); Vdd_out is the digital word controlling the power-supply voltage. The two analog signals represent respectively the VDD output generated by the power-supply, and the clock signal. The VDD was initially assigned to the nominal value of 1.1V. Considering an operating frequency of 2GHz, represented in Fig. 6, this imposes a relaxed slack time in the CP, which means that there is margin to improve power optimization (VDD reduction). Moreover, the Vdd! signal produced by the controlled power supply is being lowered, restricting power consumption when still there is margin, until it reaches values around 0.88V and 0.79V (in Fig. 7). However, if a 2.8GHz clock frequency is used (in Fig. 8), as the time slack of the CUT’s CP is near the safety margin, there is no margin left to improve power dissipation and, therefore, the Vdd! signal is not changed (note that GSO_out is the same as GAS).

Fig.6 Internal structure and port map for: (a) NOR gates, and (b) NAND gates.

Fig.7 Results for f_clock = 2 GHz.
Power simulations were also performed, to evaluate the power savings with this DVS methodology. Table I summarizes the power dissipation of the complete circuit operation for different \( V_{DD} \) values, in respect to the original circuit at nominal \( V_{DD} \) operation. As it can be seen, when comparing the original CUT without DVS and the modified CUT with DVS at nominal power-supply voltage (1.1 V), the later has a higher power dissipation (120%, i.e., 20% more), due to the DVS circuitry added. However, the purpose of the DVS methodology is to reduce \( V_{DD} \) to its minimum that still prevents errors from happening. Hence, operating with a reduced \( V_{DD} \) of 0.7 V, the power dissipation is only 66% of the original CUT without DVS circuitry, which ensures 34% of power savings, even with the added DVS circuitry. Moreover, if the comparison is made with the CUT with DVS, i.e., comparing power dissipations for different \( V_{DD} \) operation values in the complete circuit (CUT with DVS circuitry), the dissipation can reach a minimum of 55% power dissipation at 0.7 V (power is reduced by 45%), when compared to the nominal \( V_{DD} \).

### Table I. Power dissipation with DVS methodology.

<table>
<thead>
<tr>
<th>Power-supply Voltage (V)</th>
<th>Comparison with original CUT without DVS (%)</th>
<th>Comparison with CUT with DVS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>120%</td>
<td>100%</td>
</tr>
<tr>
<td>0.883</td>
<td>91%</td>
<td>75%</td>
</tr>
<tr>
<td>0.7</td>
<td>66%</td>
<td>55%</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

In this paper, a DVS methodology was presented, with additional Fault-Tolerance to delay-faults and for life-time operation. The methodology is aging-aware and controls the power-supply voltage during circuit’s lifetime to restrict power consumption. Global and local performance sensors are used to monitor on-line circuit’s timing response and both sensors complement each other on creating optimized and fail-safe circuits. DVS controller circuitry is treated as part of the CUT, ensuring error-free operation in the complete circuit (i.e., errors in the DVS circuitry are also monitored). Future work includes silicon prototyping and methodology scalability.

REFERENCES


