In-situ Fmax/Vmin tracking for energy efficiency and reliability optimization

Ivan Miro-Panades, Edith Beigné, Olivier Billoint, Yvain Thonnart

In-situ Fmax/Vmin tracking for energy efficiency and reliability optimization

Ivan Miro-Panades, Edith Beigne, Olivier Billoint, and Yvain Thonnart
Univ. Grenoble Alpes, F-38000 Grenoble, France
CEA, LETI, MINATEC Campus, F-38054 Grenoble, France
ivan.miro-panades@cea.fr

***Abstract*—Achieving the lowest possible operating voltage is needed to minimize the power consumption of a circuit but also to increase its reliability w.r.t hardware errors. An in-situ technique to estimate and reduce the design margins of a circuit is presented which significantly minimizes the operating voltage and tracks it during run-time operation of a circuit without failure. A DSP core embedding this technique has been fabricated and measured. Its V_{MIN} has been estimated within +3.5%/-2.5% at nominal clock frequency (1600MHz), thus reducing by 19% its energy per operation.**

Keywords—design margins reduction; in-situ monitors; TMFLT

I. INTRODUCTION

When the supply voltage is reduced, the power consumption of the circuit is also reduced and the reliability of the circuit is increased. Among long term critical failures modes of CMOS technology, the Time-Dependent Dielectric Breakdown TDDB) and Electromigration (EM) [1][2][3] are mostly influenced by the voltage and the temperature. Thus, reducing the voltage by 10% can lead to a failure rate reduction of 2 orders of magnitude. Furthermore, Negative Bias Temperature Instability (NBTI) increases the threshold voltage of the transistors with higher supply voltage. This in turn may eventually results in timing errors. On the other hand, there is a tradeoff between software and hardware error rates in function of the supply voltage [4]. Lowering the voltage may induce an increase the of the radiation-induced software errors, especially when operating close to threshold voltage V_{th}.

In traditional digital circuit design and to guarantee that the circuit is able to work correctly under process, voltage and temperature (PVT) variations, timing margins are added during the design. These margins constrains the circuit to a suboptimal operating point. Reducing these margins leads to a power and reliability optimization. Therefore, it is possible to reduce the voltage (V_{MIN} tracking) for an unchanged clock frequency, or to increase the frequency (F_{MAX} tracking) when the voltage is unchanged.

To reduce these margins, in-situ timing monitor techniques have been designed such as Razor II [5], Transition-Detector (TD) [6], ED latch [7], and iRazor [8]. These techniques use either an error detection mechanism or an error prevention mechanism (also named canary flip-flops). In the case of error detection [5] [7] [8], the circuit needs to detect the error and restore the previous operation to continue its process. This mechanism needs extra hardware and is not trivial to implement on a general architecture. Moreover, the detection of the error depends on the current operation of the circuit. For example, if the critical paths are on the floating-point unit of a core and that the core is in idle state, one could decide to reduce the voltage with no information from the detectors, yet creating errors elsewhere. Thus, it is not possible to guarantee that a real error occurs while the monitors don't capture it. Finally, to detect an error due to a late arrival of a long path, the hold-time margin of the circuit must be increased as it must be differentiated w.r.t. an early arrival of a short path. On the other hand, the error prevention mechanisms [6] are simpler to implement as the circuit architecture is not modified. However, the detection is based on a pessimistic detection of an error. For example, with a fixed frequency, the error will be detected some mV before the real V_{MIN} of the circuit. However, the amount of pessimism is neither known nor controllable. This pessimism must be chosen at design time of the circuit. Choosing a small pessimism increase the risk of unexpected circuit failure, while a huge pessimism will lead to sub-optimized circuit. Moreover, as for error-detection, the prevention of an error depends on the current operation of the circuit. Finally, the critical paths of a circuit are known to evolve with the supply voltage, meaning that paths that were identified to be critical at nominal voltage may no longer be critical at low voltage.

This paper present a margin reduction technique named TiMing FauLT (TMFLT) methodology that is based on in-situ monitors. It combines the advantages of error prevention techniques with a predictive circuit correlation. Thus, the unknown and uncontrollable pessimism of the monitors can be estimated and controllable. Section II provides an overview of this TMFLT methodology. Section III and IV describe the calibration phase and the run-time phase of the methodology. Section V presents the methodology to select the registers to add the monitors. Section VI concludes by summarizing the key results and insights.

II. TMFLT METHODOLOGY OVERVIEW

Unlike previous techniques where the circuit margins are estimated only during the design of the circuit, the TMFLT methodology also estimates and minimizes the actual margins in-situ. Thus the margins are computed by the circuit itself taking into account its PVT and its environment. Once the

978-1-5386-0352-9/17/$31.00 ©2017 IEEE

margins are estimated, it is possible to minimize them and track this minimum voltage point during the run-time of the circuit.

The TMFLT methodology is decomposed into two phases: the calibration phase where the circuit margins are estimated and compensated, and the run-time phase where the circuit is able to track the minimum voltage operation during run-time.

The calibration phase has to be executed at least once in the lifetime of the circuit to obtain the calibration parameters. These parameters are used by the run-time phase to track the minimum operating voltage of the circuit during normal operating mode. The calibration phase can be re-executed in order to take into account the environment change or the ageing of the circuit. Moreover, a firmware update of the circuit may improve the calibration phase and those improve the achievable margin reduction.

The TMFLT methodology is composed of two structures Timing Fault Sensor (TMFLT-S) and Timing Fault Ring (TMFLT-R). The TMFLT-S (Fig. 1a) is a timing slack sensor used to estimate the F_{MAX} / V_{MIN} of the circuit by instrumenting a given path and providing the remaining timing slack before failure. By increasing the clock frequency of the circuit, the TMFLT-S will raise an error when the timing slack is lower than a threshold. Compared to classical canary flip-flop sensor pessimism, this threshold is huge, it can be 25% of the clock period. Nevertheless, TMFLT methodology doesn't use the sensors to regulate the voltage: they are only used to build a frequency as a function of the voltage model of the circuit. Therefore, one would like to combine all TMFLT-S detected information to build a precise model and not just one.

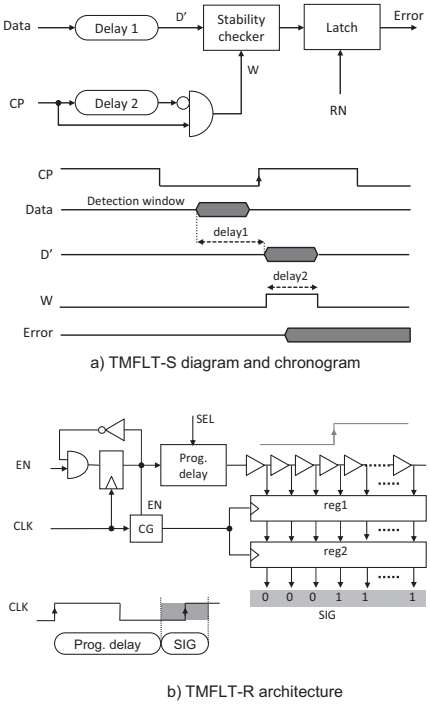

Fig. 1. TMFLT-S and TMFT-R architecture

The TMFLT-R (Fig. 1b) is a programmable replica path based on a configurable delay line able to generate warning messages when the frequency and voltage relationship are not respected (i.e. the current voltage is too low for the current clock frequency). TMFLT-S and TMFLT-R are used on the calibration phase while in the run-time phase only the TMFLT-R is used.

III. Calibration phase

The calibration phase is decomposed into two steps, in-situ characterization step where the F_{MAX} / V_{MIN} of the circuit is in-situ characterized within its environment, and TMFLT-R calibration step where the design margins are removed and calibrated into the TMFLT-R to track the minimum voltage.

The results presented in this paper have been measured on a 32bit VLIW DSP core implemented on 28nm UTBB-FDSOI technology [10].

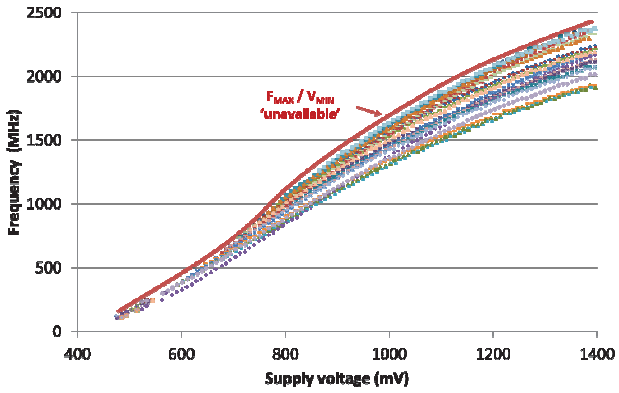

Fig. 2. TMFLT-S detection frequency as a function of the voltage (meas.)

A. In-situ characterizations step

In this step the circuit is operating in normal mode and executing a functional pattern. This pattern is a firmware code that is looped in functional mode in order to guarantee a good correlation for the characterization. For a fixed voltage, the clock frequency is set to a low frequency and progressively increased. It is also possible to have a fixed clock frequency while the voltage is set to its nominal value and progressively reduced. The information of the TMFLT-S are collected for each step. Once a TMFLT-S indicates a warning, the frequency/voltage is stored into a memory. It is possible to repeat this operation for each of the operating voltages of the circuit. Fig. 2 depicts the detection frequency of each of the TMFLT-S of the circuit for a sweep on supply voltages. Each color corresponds to a different TMFLT-S, some sensors issuing warnings before others. However, in order to know the actual margin that a given sensor provides, one should know the real maximum frequency of the circuit under test at a given voltage. This real is not measured directly die by die, as it would be too time consuming during test time and would increase the price of the circuit. It is estimated thanks to a statistical study performed on multiple dies of multiple wafers of a lot: the real is here obtained thanks to overclocking the

circuits until functional failure. On this lot, the voltage or frequency margin (frequency difference between the F_{MAX} and the detected error using the TMFLT-S) is computed for each TMFLT-S and each die. Fig. 3 shows, the voltage margin result obtained on 21 different dies for a single TMFLT-S. A statistical study on this lot is performed to compute the statistical margin of each TMFLT-S to be applied to all the circuits. This information is stored on the firmware of the circuit.

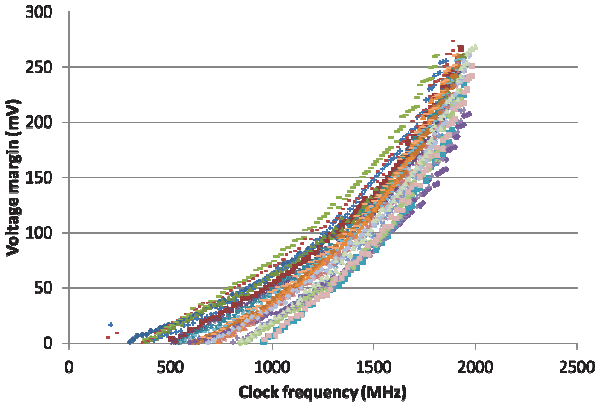

Fig. 3. Voltage margin of one TMFLT-S on 21 different dies (meas.)

Once the circuit is in the in-situ calibration step, it computes the graph Fig. 2. Then, for each TMFLT-S, the estimated chip F_{MAX} is computed from the measured detection frequency combined with its statistical margin estimation. Combining these curves by any weighing or selection operation (mean, median, percentile…) to filter measure noise, one can estimate the F_{MAX}/V_{MIN} of the circuit. Fig. 4 shows the voltage error percentage between estimated V_{MIN} voltage and real V_{MIN} voltage for each operating frequencies of the circuit using only 13 TMFLT-S. At nominal frequency, 1600MHz, this methodology reduced the voltage of the DSP core by 10% and the power consumption by 19% w.r.t worst case design. Using 21 dies, the V_{MIN} estimation error was measured to be within [-1.5%,+3%] at 700MHz (0.7V) and within [-3%,+4.5%] at 2200MHz (1.3V).

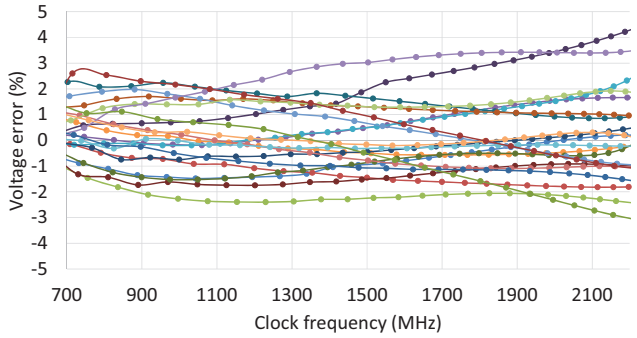

Fig. 4. Voltage error percentage between estimated and real V_{MIN} on 21 different dies (meas.)

B. TMFLT-R calibration step

Once the circuit is characterized, one can configure the TMFLT-R to issue warnings when the voltage is close to the estimated V_{MIN}. The TMFLT-R delay line delay can be configured though the SEL interface. It generates a 32bit signature (SIG), the lower the voltage the lower the signature.

Calibrating the TMFLT-R can be done by applying the estimated V_{MIN} plus a security margin into the circuit. This security margin is the combination of: TMFLT-S estimation errors on characterization, soft-error margin and environment margins. The soft-error margin may be used when the circuit is operating at low voltage to reduce the soft-error rate. Next, the functional pattern is executed in the circuit and the TMFLT-R minimum signature is collected by a controller. This value defines the minimum signature to reach for normal circuit operation. If the signature is lower, the voltage must be increased to avoid timing failures.

IV. Run-time phase

The SEL configuration and this minimum signature of the TMFLT-R are the only pieces of information needed to track the V_{MIN} during run-time.

The TMFLT controller is configured to generate a warning when the TMFLT-R signature is lower than the minimum signature. Then, the voltage is increased (or the frequency is reduced) to avoid timing violations. A proportional-integral (PI) controller is used to track the V_{MIN} of the circuit without oscillations. On the DSP core, the voltage estimation error has been measured to be within [-2.5%,+3.5%] at nominal clock frequency (1600MHz).

Comparing the TMFLT-R current signature with the minimum signature, one can estimate the amount (in mV) between the current voltage and the tracked V_{MIN}. Moreover, one can monitor the real dynamic IR-drops of the circuit. This variations depends on the circuit and on its environment (i.e. external DC/DC). Thus, it is possible to modify the configuration of the TMFLT-R in order to take into account the environment margins. Finally, the calibration phase can be replayed periodically in order to improve the calibration and to take the NBTI and ageing of the circuit into account.

V. Register selection methodology

To minimize the number of TMFLT-S added into the circuit and thus to minimize the area and power overhead, a register selection methodology has been developed. It is based on static timing analysis (STA) and back-annotated simulation of the circuit. In Fig. 5 the circuit is simulated after a first place & route trial using SDF annotation using the functional patterns. The ~50000 most critical paths according to STA are monitored during simulation, and their actual slack times are computed for the particular functional pattern. This pattern will probably not excite the most critical paths of the circuit as some registers will not be excited and other will be excited but by a path that isn't the most critical to it. Yet, the methodology only needs the registers that are excited with simple patterns and with the most critical path to them. This methodology has been used on a 900MHz multicore circuit where 256 TMFLT-S

where used per cluster [11], with a detection window of 220ps. The functional test pattern is Dhrystone benchmark. Simulation of this functional pattern allows to eliminate 41% of the 1000 most STA critical paths that where not excited and 19% that where excited but with a path that is lower than 80% of the STA delay.

After the simulation, the computed slack-time of these registers are used to select the candidates to add the TMFLT-S. Only registers excited by long paths are selected. Yet, adding a TMFLT-S on these paths may add constraints on the place & route tool (e.g. some capacitance), but as the TMFLT-S have a huge detection window, it is not necessary to select the 1% most STA critical paths of the circuit:

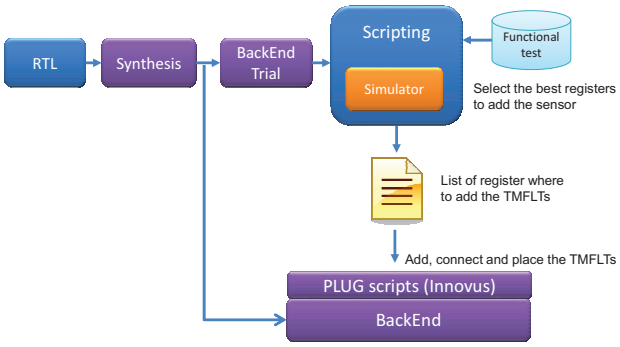

Fig. 5. Register selection methodology

Once the register list is defined, a Plug-in script is executed on the place & route tool (Innovus) to place and connect the TMFLT-S at the Pre-ClockTree stage.

After final place & route, the circuit is again simulated with the functional pattern for verification. Fig. 6 depicts the probability of detecting a TMFLT-S before reaching the F_{MAX}. The horizontal axis represents the clock period margin before F_{MAX}. Thanks to selecting the appropriate registers for this particular pattern, only 2% of the TMFLT-S will fail detecting before the F_{MAX}, and 88% of the TMFLT-S detect within one-half (110ps) of the detection window, showing accurate register selection. If the paths where only selected based on STA, these number may be very low as more than 60% (41% + 19%) of the 1000 STA most critical paths are useless.

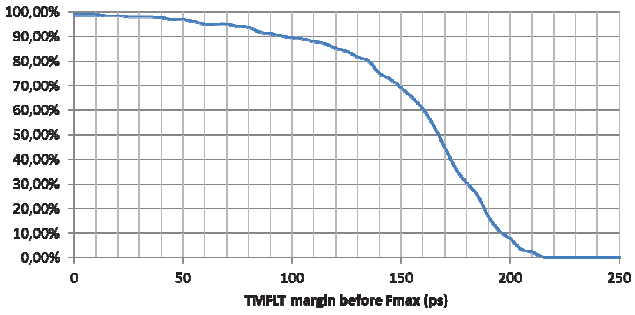

Fig. 6. Probability of detecting the F_{MAX} with a clock margin

VI. Conclusions

An efficient way to increase both the energy efficiency of a circuit and its reliability is to reduce its operating voltage. The TMFLT technique to minimize the operating voltage without generating a timing failure allows high performance both at nominal voltage and low voltage. The TMFLT technique estimates in-situ the voltage margin of the circuit and reduces it to the minimum. The amount of residual margin can be tuned after circuit fabrication in order to take the circuit environment and the IR-drop into account. Moreover, at low voltage it is possible to increase the margin to mitigate the soft-error rate. A DSP was fabricated on UTBB-FDSOI technology using this technique to estimate and track the V_{MIN} of the circuit where its power consumption was reduced by 19% at nominal frequency. A register selection methodology has been employed on a multicore design where 98% of the inserted monitors will issue a warning before the F_{MAX}. Thanks to dynamic adaptation and recalibration of the circuit V_{MIN} using the proposed monitors, failure rate can be reduced by orders of magnitude.

References

[1] X. Li, J. Qin and J. B. Bernstein, "Compact Modeling of MOSFET Wearout Mechanisms for Circuit-Reliability Simulation," in IEEE Transactions on Device and Materials Reliability, vol. 8, no. 1, pp. 98-121, March 2008.

[2] S. Kosonocky, T. Burd, K. Kasprak, R. Schultz and R. Stephany, "Designing in scaled technologies: 32 nm and beyond," 2012 Symp. on VLSI Technology (VLSIT), Honolulu, HI, 2012, pp. 147-148.

[3] J. R. Black, "Electromigration—A brief survey and some recent results," in IEEE Transactions on Electron Devices, vol. 16, no. 4, pp. 338-347, Apr 1969.

[4] K. Swaminathan et al., "BRAVO: Balanced Reliability-Aware Voltage Optimization," 2017 IEEE International Symposium on High Performance Computer Architecture, Feb. 2017.

[5] S. Das et al., "RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance," in IEEE Journal of Solid-State Circuits, vol. 44, no. 1, pp. 32-48, Jan. 2009.

[6] K. A. Bowman et al., "Energy-Efficient and Metastability-Immune Resilient Circuits for Dynamic Variation Tolerance," in IEEE Journal of Solid-State Circuits, vol. 44, no. 1, pp. 49-63, Jan. 2009.

[7] S. Kim and M. Seok, "Variation-Tolerant, Ultra-Low-Voltage Microprocessor With a Low-Overhead, Within-a-Cycle In-Situ Timing-Error Detection and Correction Technique," in IEEE Journal of Solid-State Circuits, vol. 50, no. 6, pp. 1478-1490, June 2015.

[8] Y. Zhang et al., "8.8 iRazor: 3-transistor current-based error detection and correction in an ARM Cortex-R4 processor," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2016, pp. 160-162.

[9] X. Li, J. Qin and J. B. Bernstein, "Compact Modeling of MOSFET Wearout Mechanisms for Circuit-Reliability Simulation," in IEEE Transactions on Device and Materials Reliability, vol. 8, no. 1, pp. 98-121, March 2008.

[10] E. Beigné et al., "A 460 MHz at 397 mV, 2.6 GHz at 1.3 V, 32 bits VLIW DSP Embedding F MAX Tracking," in IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 125-136, Jan. 2015.

[11] P. Vivet, C. Bernard, E. Guthmuller, I. Miro-Panades, Y. Thonnart and F. Clermidy, "Interconnect Challenges for 3D Multi-cores: From 3D Network-on-Chip to Cache Interconnects," 2015 IEEE Computer Society Annual Symposium on VLSI, Montpellier, 2015, pp. 615-620.