Power and Performance Optimization in Long-term Operation

A. Romão, J. Semião, C. Leong, M. B. Santos, I. C. Teixeira, J. P. Teixeira

Abstract The work developed consists in a power or performance optimization methodology, for long-term operation, using global and local aging aware performance sensors. Methodology allows circuits to be dynamically optimized, during their life-time, according with one of two possible needs: (1) restrict power consumption, by reducing power-supply voltage to the minimum value that prevents errors from happening; or (2) optimize performance, by increasing operating frequency to the maximum limit that prevents errors' occurrence. The dynamic optimization is achieved by using a cooperative work of global and local sensors. Moreover, new global sensor architecture, controller and a DCO (Digital Control Oscillator) are presented, to demonstrate frequency automatic optimization, according with sensors' outputs. Extensive spice simulations in a 65nm CMOS technology demonstrate the results.

Keywords Performance sensor, Dynamic Voltage Frequency Scaling (DVFS), Frequency and Power optimization, PVTA variations.

1 Introduction

CMOS circuit's performance is affected by many parametric variations, like Process, power-supply Voltage and Temperature (PVT) [1], as well as to aging effects (PVT and Aging – PVTA), with the Bias Temperature Instability (BTI) being one of the main concerns that affects transistors. These variations, regardless of their origin, may lead to chip failures.

This work is a continuation of the work reported in [11], and the original contributions are: (i) an improved global sensor (GS) architecture, to include 2 dummy critical paths (CP) and sensors, targeting NBTI (Negative BTI) and PBTI (Positive BTI) degradations; (ii) the development of a new Control Module for on-line optimization of power or performance efficiency; (iii) methodology application to a complete case-study for on-line performance efficiency optimization, including a Digitally Controlled Oscillator (DCO) to allow circuit's clock frequency variations. GS are used to perform periodic on-line monitoring. LS are used to keep GS tuned with circuit's performance. Both are sensitive to PVTA (Process, power supply Voltage, Temperature and Aging) variations and both can trigger the supply Voltage or Frequency on-line tuning.

2 Improved Global Sensor Architecture and Control Mechanism

Fig. 1 presents the new GS architecture, with two CP delay's replicas implemented with NOR and NAND gates. The difference from the global sensor (GS) described in [11] relies on the 2nd dummy CP created using NAND gates. These two dummy paths must have higher propagation delays than the expected circuit's CP. The NOR gates' chain will have, presumably, a higher aging degradation when compared with circuit's CP if NBTI degradations are considered, while the NAND gates' chain will have, presumably, a higher aging degradation when compared with circuit's CP if PBTI degradations are

This work was partially supported by EU COST Action IC1103 – MEDIAN – "ManufacturablE and Dependable multIcore Architectures at Nanoscale".

A. Romão and J. Semião are with Department of Electrical Engineeting, ISE - University of Algarve, Faro, Portugal {a15297, jsemiao}@ualg.pt
M. B. Santos, I. C. Teixeira and J. P. Teixeira are with IST, Techncal University of Lisbon, Lisboa, Portugal, {marcelino.santos, isabel.teixeira, paulo.teixeira}@ist.utl.pt
C. Leong is with the INESC-ID Lisboa, Lisboa, Portugal, carlos.leong@inesc-id.pt.

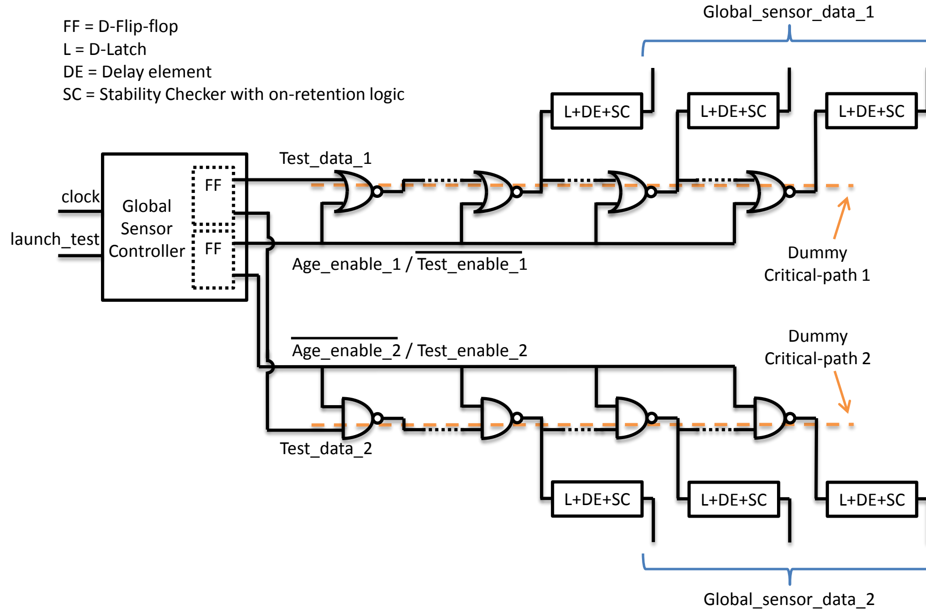


Fig. 1 Global Aging-Aware Performance Sensor architecture.

considered. When the local control unit in the GS receives a signal to start the analysis control and data signals are generated to produce the two state transitions (Low-High and High-Low) in the dummy paths. Moreover, sensor cells are connected at the output of several NAND and NOR gates along the paths, to create several fictitious paths with different propagation times. The architecture of these sensor cells was defined in [11], and also more details can be found for the NOR gate internal structure and port-map, to achieve a high aging degradation in the fictitious paths (a similar analysis can be drawn for the NAND gates).

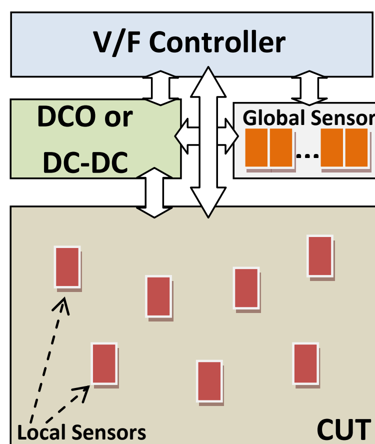
The V/F optimization can be achieved with a reduced error margin, defined by sensors' architecture. However, a calibration procedure is mandatory in test mode, to tune GS and LS. This procedure, uses off-line tests (e.g., scan-based delay-fault oriented tests), performed repeatedly for different values of the clock frequency (or voltage), and obtains the maximum (or minimum) frequency (or voltage) for which no LS is activated (with a controlled error margin defined by design). This information will be stored in a register (V/F register) and corresponds to the nominal frequency (or voltage). Furthermore, the GS output for this nominal frequency (or voltage) also defines the safe/optimized output for GS, and this information is stored (in GSO_{safe} registers) and will be used during the on-field operation.

In an on-field 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 frequency (or VDD voltage) adjustments. The GS activation period is defined according with available time slack (the shorter the time slack, the shorter the GS's activation should be).

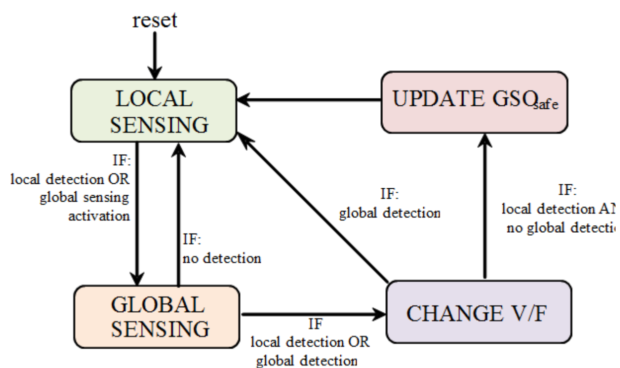
Fig. 2 depicts the block diagram of the circuit with sensors and controller, and the simplified functionality of the power or frequency optimization controller. As LS are always on-line, the reset state is "Local Sensing". "Global Sensing" is triggered when an error is detected by LS, or by a timer. 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/F" state and the frequency (or voltage) 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 GSO_{safe}" state.

3 Methodology Demonstration and Results

For sensors' results please refer to [6][11][12]. In this paper the purpose is to present complete system simulation results. Simulations were performed with HSPICE for the



(a)



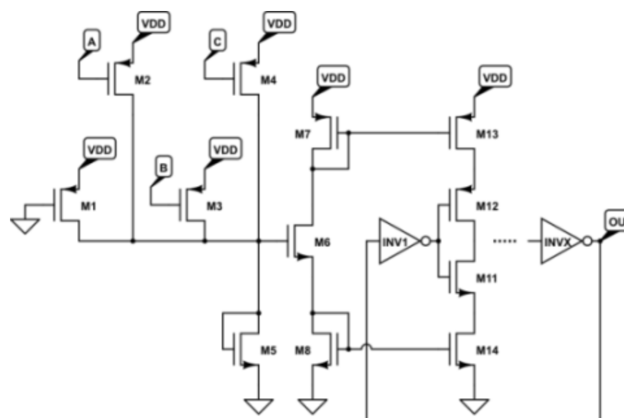
(b)

Fig. 2 (a) Block diagram; (b) Simplified functionality for the V/F controller.

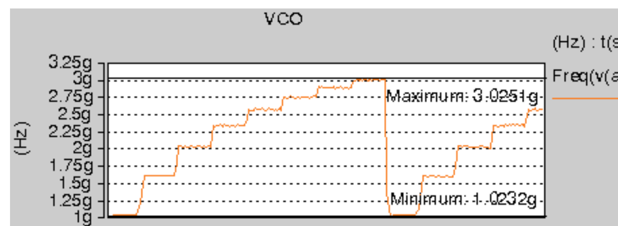
CUT (PM4-2), a 4-bit, 2-stage Pipeline Multiplier with 9 inputs (8 data and 1 clock), 8 data outputs in PTM 65nm [7], with a nominal V_{DD} of 1.1V value and ±10% variations for worst case conditions, with 27ºC for nominal temperature and 110ºC for severe aging degradations.

For complete methodology demonstration, a frequency optimization procedure was implemented. A DCO block was defined, as described in Fig. 3(a), which comprises a ring oscillator with 5 stages, combined with a bias and control circuitry. In order to generate 8 different frequencies (with a 3-word data input bus, ABC), the M2, M3 and M4 transistors' sizes must be obtained by simulation, achieving a frequency range from 1GHz to 3GHz (Fig. 3(b)). Moreover, a 3-word data was also used to code the controller registers, for sensors' safe output and for storing the nominal frequency/VDD. 8 GS outputs for each dummy CP were used, and 3 LS were inserted in the CUT, with their output connected through an OR gate to the control module.

SPICE simulations were performed for the complete circuit and the results are summarized in Fig. 4. VFcode and GSOcode (identified as safe/optimized frequency and corresponding GS output data) were set to "100" and "011" respectively, which is purposely a non-optimized and lower frequency, and the "100" initial VFcode imposed a frequency of 2.2GHz in the DCO output. In Fig. 4, the first graph summarizes the buses in decimal value, for the controller state, VFcode, GSOcode (the safe code for GS) and GSOread (GS output, only for the dummy CP #1). The second graph presents the simulated V_{DD} value, with a modulation in this signal by changing temporarily V_{DD} to 90% of its nominal value, to emulate the existence of a PVTA variation. The third graph presents the clock signal in two forms, time and frequency domains, while the fourth graph presents again the clock in the frequency domain, and the LS signal. Note that initially, as there are no local errors and the reading of the GS (GSOread) is lower than the GSOcode, the Controller raises frequency (VFcode is lowered) until GSOread equals the GSOcode. However, when V_{DD} is lowered, local errors start to be signalized by the LS (the error didn't actually occurred), and the VF Controller reduces frequency (by raising VFcode) (4th graph of Fig. 4).



(a)



(b)

Fig. 3 (a) DCO architecture; (b) DCO frequencies for the 8 ABC input combinations.

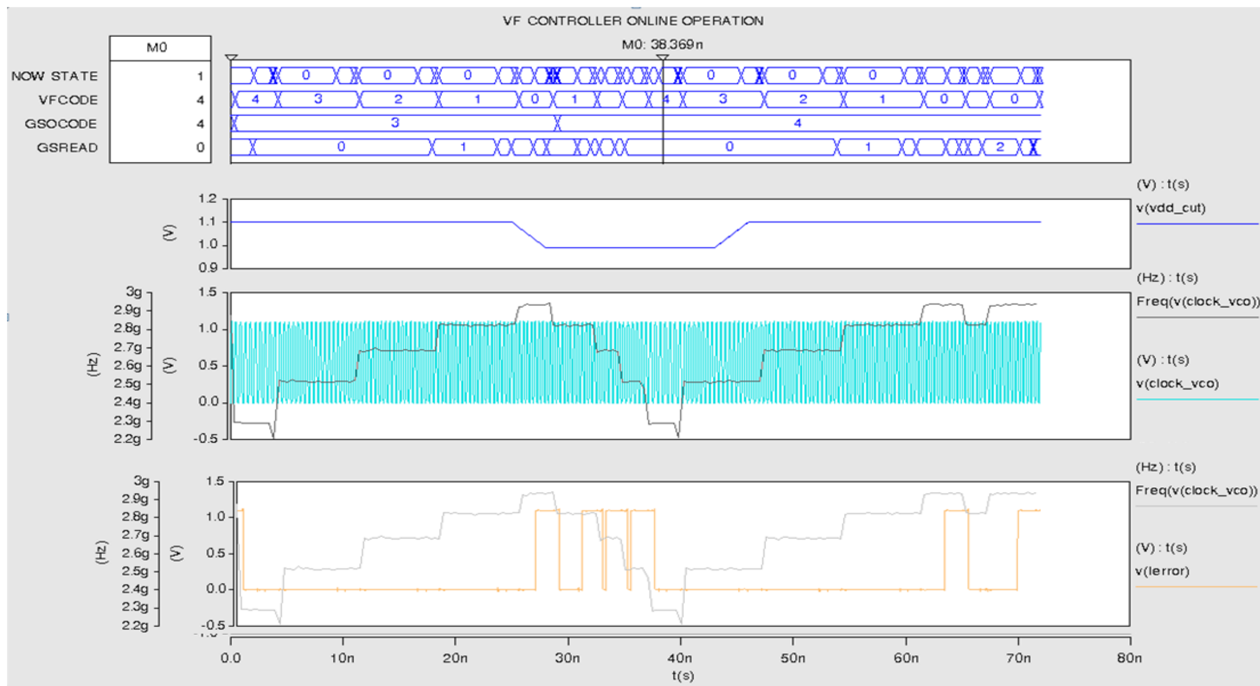


Fig. 4 HSPICE simulation results for complete circuit operation.

References

1. J. Semião, et al., "Signal Integrity Enhancement in Digital Circuits", IEEE Design and Test of Computers, September-October 2008, vol. 25, no. 5, pp. 452-461
2. D. Kim, J. Kim, M. Kim, J. Moulic, H. Song, "System and Method for Monitoring Reliability of a Digital System", IBM Corp., US Patent 7495519, Feb. 24, 2009.
3. M. Agarwal, et al., "Circuit Failure Prediction and Its Application to Transistor Aging". Proc. VLSI Test Symp. (VTS), pp. 277-286, 2007.
4. J.C. Vazquez, et al., "Predictive Error Detection by On-line Aging Monitoring", Proc. IEEE Int. On-Line Test Symp. (IOLTS), 2010.
5. C. Martins, et al., "Adaptive Error Prediction Flip-Flop for Performance Failure Prediction with Aging Sensors", Proc. of IEEE VLSI Test Symposium (VTS), April 2011.
6. C. Martins, et al., "Adaptive Error-Prediction Aging Sensor for On-Line Monitoring of Performance Errors", Proceedings of the 26th Conference on Design of Circuits and Integrated Systems - DCIS'2011, November, 2011.
7. Predictive Technology Model (PTM), http://www.eas.asu.edu/~ptm/.
8. Hans Jacobson, et al., "Stretching the Limits of Clock-Gating Efficiency in Server-Class Processors", 11th International Symposium on High-Performance Computer Architecture (HPCA'05), San Francisco, California, February, 2005, ISBN: 0-7695-2275-0.
9. Kaijian Shi and David Howard, "Sleep Transistor Design and Implementation - Simple Concepts Yet Challenges To Be Optimum", Proc.. IEEE VLSI-DAT, April, 2006.
10. D. Blaauw, et al., "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance," Solid-State Circuits Conference. ISSCC 2008. Digest of Technical Papers. IEEE International , vol., no., pp.400-622, 3-7 Feb. 2008.
11. J. Semiao, et al., "Aging-aware Power or Frequency Tuning with Predictive Fault Detection", IEEE Design & Test of Computers, Volume 29 , Issue 5, September/October 2012, DOI: 10.1109/MDT.2012.2206009.
12. J. Semiao, et al., "The Influence of Clock-Gating On NBTI-Induced Delay Degradation", 18th IEEE International On-Line Testing Symposium - IOLTS'12, Sitges, Spain, June 27-29, 2012.