Timing in-situ Monitors: Implementation Strategy and Applications Results

A. Benhassain[1,2], F. Cacho[1], V. Huard[1], M.Saliva[1,2] , L. Anghel[2] ,C. Parthasarathy[1], A. Jain[1], F. Giner[1],

[1]STMicroelectronics, Technology R&D, Crolles, France
Phone: + 33(0)438922536, e-mail : sidi-ahmed.benhassain@st.com
[2]TIMA 46, avenue Félix Viallet, 38031 Grenoble, France

***Abstract* – In-situ monitor is a promising strategy to measure timing slacks and to provide pre-error warning prior to any timing violation. In this work, we demonstrate that the usage of in-situ monitors with a feedback loop of voltage regulation is suitable for process compensation, multiples OPP modes, temperature and ageing compensation.**

***Index Terms* — in-situ timing monitors, CMOS reliability, timing margin.**

I. INTRODUCTION

With CMOS technology scaling, it becomes more and more difficult to guarantee circuit functionality for all process voltage temperature (PVT) corners. Moreover, circuit wear-out degradation lead to additional temporal variation, it results in an increase of design margin for reliable systems [1]. Adding pessimistic timing margin to guarantee all operating conditions under worse case conditions is no more acceptable due to the huge impact on design costs.

One can report two categories of ageing monitoring techniques. Firstly, we can define standalone sensors utilizing various configurations of ring oscillators [2] and delay chain. Replica paths [3] are a solution to mimic the timing behavior of the original path in combinatory logic. Second, in-situ delay monitors can directly measure the delay degradation of a specific path within the target circuit, this approach is very promising to provide reliable timing information [4]. Delay monitors such as “Razor I” [5] and “Razor II” [6] detect timing errors in actual paths and a procedure a local micro-rollback execution ensuring error correction. The Adaptive Voltage Scaling (AVS) approach in [7, 8] proposes error correction by using in-situ monitors able to detect timing error and global system action following the error detection. Another approach consists in detecting timing pre-error instead of timing error by detecting critical transitions [8]. In this case, the in-situ delay monitors can be used as a reliability technique to provide alert prior setup violation. This technique is also further combined with global system actions such as AVS or DVFS.

In this paper, an innovative insertion flow of monitor is presented. Two solutions of ISM are discussed and compared. The first one is built with standard cells available in the technology design platform library, named here built-in flow ISM. The second one uses a dedicated custom design, named cell-based ISM. Finally, several applications of ISM usage for compensation are presented.

II. ISM INSERTION FLOW

The advantage of ISM located inside digital block is the capability to accurately capture all sources of local physical, environmental and temporal variations. ISMs under investigation are presented in the Fig. 1. The basic idea is to delay the data of a critical path arriving at D in the shadow FF, and to compare it with the regular FF. When pre-error signal rises, it means that a violation of the setup time has occurred in the shadow FF and the remaining slack of the data path is close to the timing of the delay element, as defined in the schematic. In our scheme, pre-error signal can suffer from shadow FF metastability. However, unlike Razor I which have an instantaneous detection and one clock cycle later correction, here a decision of feedback loop regulation based on pre-errror is taken after a large amount of ‘1’ pre-error signal. Thus, a dedicated metastability detector is not required.

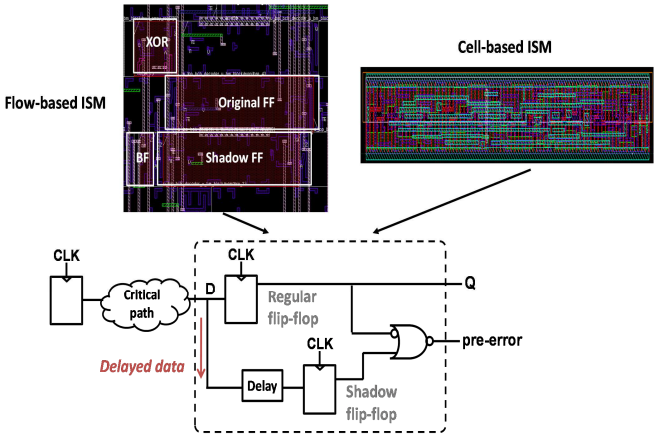

Fig. 1. Schematic and layout of in-situ monitor under investigations. Data arriving at Q is delayed in shadow FF and compared to the regular one. Flow-based ISM is composed of standard cells available in the design platform. Cell-based ISM is a fully customized design.

This schematic can be carried out in two different ways: semi-custom (flow-based ISM) or full custom designs (cell-based ISM). In the first one, all schematic elements are issued from the standard cell design platform. Placement and connectivity is performed with scripting during the flow execution. The second one is a new cell dedicated this usage, fully optimized where all parameters are fixed and controllable. For that approach, all CAD views of the new cell (functional, physical, delay, power, etc.) need to be developed to be compliant with standard digital flow.

The generic approach is illustrated in the Fig. 2. The classical Front-end steps are executed with synthesis and floorplaning. At the end, a gate netlist is provided as input to placement and route tool. After placement and pre clock tree synthesis (CTS), a timing analysis (TA) is performed. For setup functional corner, a decision is made to insert monitor (FF cell sweep for cell-based ISM) and to regenerate connectivity on a sub-set of critical path. It results in a new gate netlist, new timing and power figures, and the flow is normally re-executed: post CTS (hold and setup optimization), route and optimization until the design is timing, power and reliability closed. A certain number of back and forth steps is required to fully satisfy the initial design specification, as shown in Fig. 2.

For illustration, some timing analyses are presented in the Fig. 3. Based on an initial 5% worst slack selections, ISM are inserted in a sub-set of path. Note that another strategy can be used, Lai [9] select and monitor paths whose worst-case delay exceeds the typical operating clock period. At step 3 (Fig. 2), histogram of paths are reported for an implementation with and without ISM. In the following analysis, delayed paths are not reported.

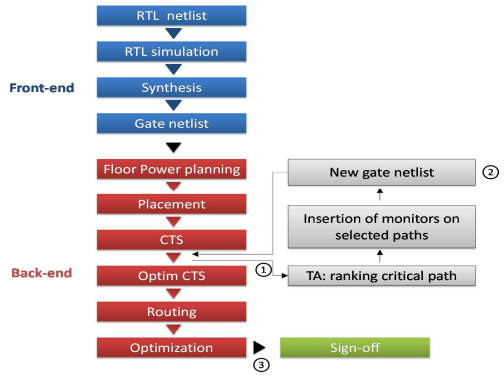

Fig. 2. Flow insertion of in-situ built-in monitors. During Front-End flow, a preliminary Timing Analysis is performed after pre-CTS step. In-situ monitors are inserted in sub-set of critical paths and a new gate netlist is generated. Then the Back-end flow is normally executed with new gate netlist.

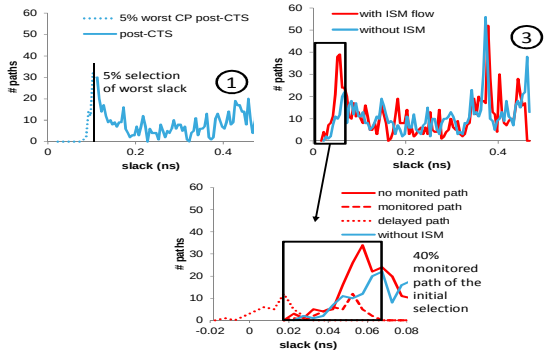

Fig. 3. Timing analysis of BCH results at different step of the flow. A preliminary TA at post-CTS is calculated (step 1). Based on this ranking, 5% worst slack are selected, and ISM are inserted. In the final TA (step 3), slack of monitored and none monitored paths are presented.

A particular attention is paid to be sure that for flow-based ISM the inserted cells are physically the closest possible to the monitored FF. To achieve this objective, timing constraints are adapted to minimize the skew between shadow and regular FF. Moreover the delayed data arriving at shadow FF is not considered as a real path when the place and route tool optimize to fulfill the timing constraint. It means that there is theoretically no timing penalty after ISM insertion expected the one induced by the slightly additional routing resource.
Thus, the distribution of the CP and sub-CP histogram is important to analyze when using this approach. The violation hazards of a path due to the induced ageing failures are a function of its remaining slack. However, for the ISM flow, we use the number of inserted ISM as the metric under investigation.

III. Example of Applications

After discussing the strategy of insertion ISM flow, some experimental results are now reviewed. Dedicated digital block are developed where on 10% of critical paths custom cell-based ISM has been inserted. We have investigated designs in 28nm, Low Power (LP) and Fully Depleted SOI (FDSOI) developed at STMicroelectronics. Digital block studied is Bose, Ray-Choudhary and Hocquenghem (BCH) error correcting code IP consisting of encoder and decoder modules. More details about this circuit can be found in [1].

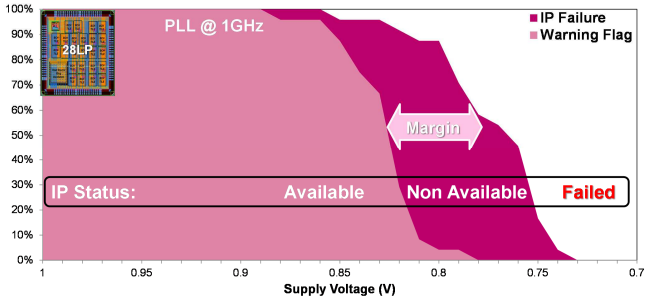

Fig. 4. Management of multicore architecture using ISM. At fixed 1GHz clock, when decreasing supply voltage, a warning flag appears earlier before the IP failure. The 18 core safety margins are in a 100mV supply voltage range.

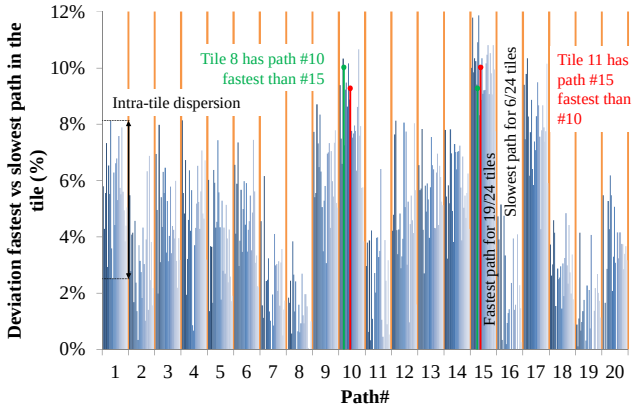

Fig. 5. Illustration of intra-tile dispersion for 20 logic paths. The slowest path in the tile is normalized at 0%. The most critical path may vary in function of the tile.

In the first application ISM are inserted in the architecture to manage the variability under optimum power budget. Major challenge in multicore architecture is to cope with inter-core

dispersion. Indeed, local process dispersion leads to variation of speed and thus power consumption of all cores. To tackle this dispersion, an additional margin in the voltage stack needs to be used. It is not a trivila task to establish this margin because it is deeply influenced by the process centering and dispersion of manufacturability. Alternative approach is to insert ISM and to use their flag as a warning to be considered as inputs of margin capabilities. As depicted in Fig. 4, 18 BCH cores are implemented in LP technology with ISM without any feedback loop. Under constant 1GHz clock frequency, when supply voltage is decreased, a first flag monitor occurs at 0.99V, corresponding to a 1% of voltage decrease. At that point, the operating functionality is still correct. While supply voltage continue to decrease, more and more flags occur on different cores and a first failure is reported (setup violation) at 0.85V. Interestingly, the V_{MIN} (minimal voltage sustaining to maintain functionality at a given PLL clock) distribution for all 18 cores, depends on the application execution of all cores and their ageing experience. To optimize the choice of voltage stack in multicore architecture, the strategy would be to monitor the first flag of each core instead of using a conservative extra margin covering intra-core dispersion. Fig. 5 shows dispersion of 20 delay paths in the 18 core.

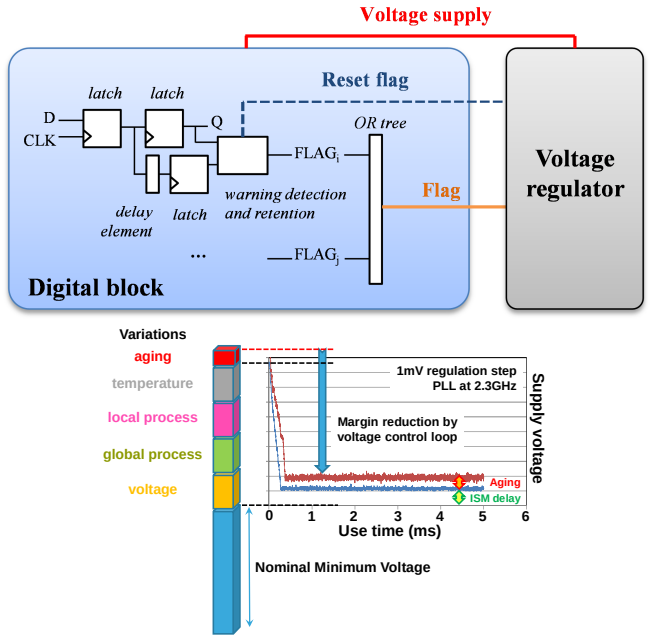

Fig. 6: Voltage regulation scheme. A regulation voltage is performed after a signal of ISM flag. Illustration of voltage regulation results: with a 1mV regulation step, the voltage is dynamically decreased until a flag signal is received. Once a flag signal is found, voltage is increase by step. Strategy is relevant to account for ageing in time.

The second application focuses on voltage regulation. The principle of concept is explained in Fig. 6, and consists in regulating the voltage in function of the flag tree issued. In following results, regulation feedback loop is carried out by hardware equipment with a fast time of enslavement. Proof of concept of regulation driven by ISM is now described. Basically, the voltage stack is composed of all existing margin (voltage range, global process, local process, temperature range variation and ageing. It covered all operating conditions, process variation and deviation in time. A strategy of ISM enables to suppress all fixed margins and to give only the needed one to guaranty the functionality. Illustration is presented in Fig. 6 bottom. Starting from a reference voltage (as given by voltage stack), the voltage is dynamically decreased until a first flag warning rises. At the first flag, regulation increases the voltage by a step. A control loop does the dynamical regulation at each millisecond. As demonstrated in Fig. 7, this scheme is suitable for ageing, indeed after ageing, logic cells are slowdown, an additional supply is needed compared to fresh situation to sustain the same clock period. The same regulation (red line) is performed for aged circuit.

A formalism describing the fast voltage change during regulation was described in [1] with Markov Chain. The assumption here is that during execution of the application, logic path are used in a random manner. Thus, monitor flags can rise respecting a Gaussian distribution in time. It results in a Gaussian distribution of the increase-decrease step in time during regulation. It is noticeable that this feature is greatly application dependent, indeed an application can exercise repetitively limited amount of logic paths.

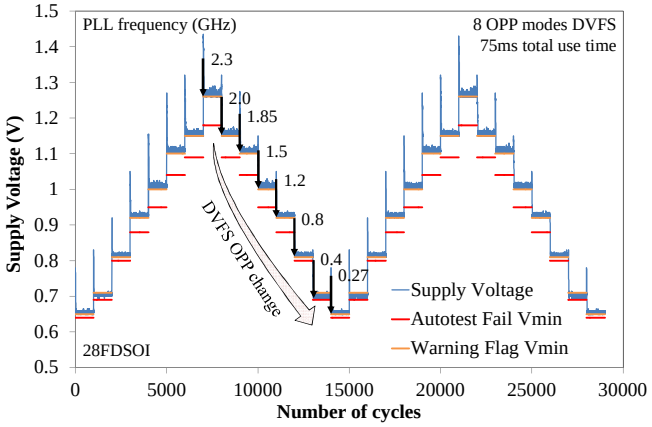

Fig. 7: Support of 8 OPP modes using voltage regulation scheme. Use time of each mode is 75ms. Each mode is carried out at one condition of frequency, voltage is automatically adjusted according to flag warning output.

Voltage regulation scheme is a promising strategy to cope with all source of variation, fast (process, temperature, voltage) and long-term (ageing). Generalization and usage extension of this scheme is now presented. Dynamical Voltage Frequency Scaling is now widely used in digital conception, it consists in adaptive a point of voltage operation (AVS) in function of usage requirement or Operating Performance Point (OPP) which is typically the frequency range. For example a low resource consuming program can be executed at a lower voltage than the required voltage to execute a high performance application (signal processing…). The conception of digital block with OPP is more difficult to handle, the sign-off must be done for a large amount of operating conditions: multi-mode and multi-PVT. ISM demonstrates a great interest for the OPP capability as explained in Fig. 7. To mimic different OPP modes, different clock frequency are applied to the block from 2.3GHz to

0.27GHz. For a given mode, supply voltage is decreased until a first flag occurs, regulation is performed in order the never find flag at minimal supply. Then, mode changes (by frequency), and supply is automatically adjusted to the lower voltage. The waveform of OPP modes is repeated in increasing and decreasing order.

The yellow line represents the boundary provided by flag warning. Note that voltage supply dynamically cross this boundary to reach its minimal value. Red boundary is the limit of functional failure. As it is reported in Fig. 5, the maximal benefit of power, supply voltage provided by ISM minus supply at which IP fails, is at low voltage. At high speed the benefit is lower than at lower frequency.

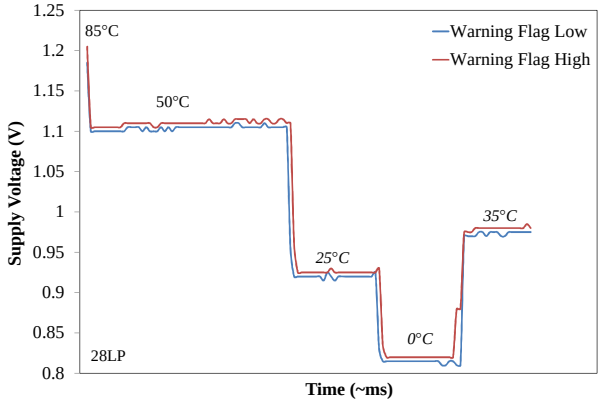

Fig. 8: Support of temperature change using voltage regulation scheme. External temperature change results in a dynamical adaptive supply voltage according to flag signal output.

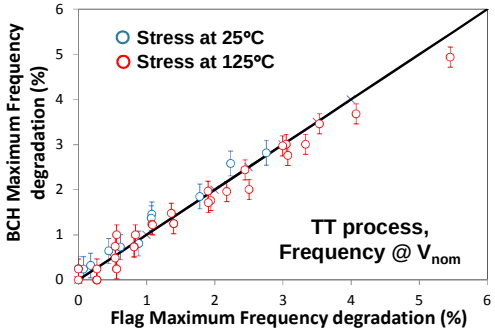

Fig. 9: Correlation between the ageing of circuit and the ageing ISM at 25°C and 125°C.

Another aspect is the temperature variation. The scheme of voltage regulation is functional under abrupt temperature change. Fig. 8 shows an external variation of temperature imposed by tester, and the evolution of adaptive supply voltage. For a constant clock period, a fast temperature decrease, the supply is dynamically decreases. It follows the trend of silicon performance in function of temperature.

Finally, concerning wear-out mechanism, ISM is an excellent monitor to describe the ageing of digital block. Fig. 9 highlights the fact that during stress at V_{MAX} biasing, the degradation of IP failure F_{MAX} is correlated to the flag F_{MAX}. A stress at 25°C and 125°C preferentially exacerbated the Hot Carrier Injection and the Bias Temperature Instability. The major conclusion drawn is that fresh critical logic path ranking is quite similar also after ageing. It can be explained by a pseudo-randomization of path usage and thus a homogenous ageing. Indeed, input vectors are internally generated by a pseudo-random binary sequence.

The interest of ISM usage with a voltage regulation is demonstrated. This scheme is suitable for process compensation, multiples OPP modes, temperature and ageing compensation. It worth noticing that, voltage biasing regulation was selected for LP technologies, but back biasing regulation is preferred for FDSOI. It has the advantages to have a better power reliability trade-off as recently reported [10].

V. Conclusion

A flow of in-situ monitor is developed and applied to different circuits. Two types of monitors are compared and discussed: cell-based and flow-based approach. Performance penalty and area overhead of ISM is slightly small. Some applications of adaptive regulation are illustrated, this scheme is promising for process compensation, multiples OPP modes management, temperature (voltage dynamically decreases with temperature) and ageing compensation (F_{MAX} of ISM flag and F_{MAX} of IP are correlated and follow the same degradation under stress.

Acknowledgments

The authors would like to thank the overall wafer level reliability, design and test teams for their outstanding support.

References

[1] V. Huard, “Adaptative wear out management with in-situ management” IRPS 2014

[2] X.Wang ,”Path-RO: a novel on-chip critical path delay measurement under process variation IEEE ACM(2008)

[3] S.Wang “Representative Critical Reliability Paths for low-cost and accurate on-chip aging evaluation” IEEE/ICCAD (2012)

[4] Saliva.M ,”Digial circuits reliability with in-situ monitors in 28nm fully depleted SOI “ IEEE/DATE (2015)

[5] S. Das et al.,”A Self-Tuning DVS Processor Using Delay-Error Detection and Correction” IEEE J. Solid-State Circuits, Apr. (2006)

[6] D. Blaauw et al.,”RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance” IEEE J. Solid-State Circuits,Jan. (2009)

[7] K.A.Bowman “Energy-efficient and Metastability-Immune Resilient Circuits for Dynamic Variation Tolerance” IEEE Journal of Solid-State Circuits

[8] M. Wirnshofer “A Variation-Aware Adaptive Voltage Scaling Technique Based on In-situ Delay monitoring” IEEE/DDECS (2012)

[9] L.Lai, et. al. “SlackProbe: A flexible and efficient in situ timing slack monitoring methodology,” T-CAD, vol. 33, no. 8, (2014)

[10] P. Mora .” 28nm UTBB FDSOI product Reliability/Performance trade-off optimization through body bias operation”, IPRS2015