Timing Slack Monitoring under Process and Environmental Variations: Application to a DSP Performance Optimization

Philippe Maurine, Bettina Rebaud, Marc Belleville, Edith Beigné, Christian Bernard, Michel Robert, Nadine Azemard

Timing slack monitoring under process and environmental variations: Application to a DSP performance optimization

B. Rebaud [a], M. Belleville [a], E. Beigné [a], C. Bernard [a], M. Robert [b], P. Maurine [b,*], N. Azemard [b]

[a] CEA, LETI, MINATEC campus, F38054 Grenoble Cedex, France
[b] LIRMM, Montpellier, France

Keywords:

Variability
Monitor
Timing slack
Process compensation

A B S T R A C T

To compensate the variability effects in advanced technologies, Process, Voltage, Temperature (PVT) monitors are mandatory to use Adaptive Voltage Scaling (AVS) or Adaptive Body Biasing (ABB) techniques. This paper describes a new monitoring system, allowing failure anticipation in real-time, looking at the timing slack of a pre-defined set of observable flip–flops. This system is made of dedicated sensor structures located near monitored flip–flop, coupled with a specific timing detection window generator, embedded within the clock-tree. Validation and performances simulated in a 45 nm low power technology, demonstrate a scalable, low power and low area system, and its compatibility with a standard CAD flow. Gains between an AVFS scheme based on those structures and a standard DVFS are given for a 32 bits VLIW DSP.

1. Introduction

If the scaling of CMOS technologies has brought amazing integration capabilities, it has also recently led to a dramatic design margin increase [1] mainly explained by the increase of process variations and the conservatism of worst case analyses and CAD tools. To cope with this design margin increase, statistical techniques [2–4] have been identified as a mean to obtain better estimations and thus as a solution to reduce design margins, while maintaining a good yield.

Unfortunately statistical techniques may not be applied to take dynamic variations into account. As a result, dynamic variations like voltage and temperature variations [5,6] or ageing effects [7] are still taken into account considering worst case situations. To get around this limitation and further reduce design margins, two different approaches may be adopted.

The first one consists in integrating specific structures or sensors to monitor in real-time the physical and electrical parameters required to control dynamically the operating frequency and/or the supply voltage and/or the substrate biasing. Several process voltage temperature (PVT) sensors have been proposed n the literature [8–12] for global variability compensation. However, the use of such PVT sensors has some limitations.

Firstly their area and power consumption may be high and thus their number has to be limited. Secondly their use requires the integration of complex control functions in LUT and intensive characterizations of the chip behavior w.r.t. the considered PVT variables. Finally the use of Ring Oscillator (RO) structures [8,9,12] to monitor the circuit speed can be another limitation, since an RO may have sensitivities to PVT quite different than the behavior of real circuit datapaths. However, this can be partly overcome in adopting a replica path approach as proposed in [13]. It consists in monitoring the speed of some critical paths that are duplicated in the sensors to replace the traditional RO.

The second approach, to compensate PVT variations and ageing effects, is to monitor directly the sampling elements of the chip (latches or D-type flip–flop) to detect the occurrence of timing faults. This can be achieved either by inserting specific structures or using ad-hoc sampling elements [14–17] detecting the occurrence of a timing violation, by performing a delayed comparison or by detecting a signal transition within a given time window. The main advantage of this approach is its ability to detect the effects on timings of local and dynamic variations occurring in the close vicinity of the inserted specific structures. A second and significant advantage is the simple and binary sensor output.

However, this second approach has also some drawbacks. Indeed, an important number of sensors might be required to obtain a full coverage of the circuit. Thus, these structures must be as small as possible and consume a small energy when the circuit operates correctly. In addition, the detection of an error

requires the full replay of the faulty data processing at a lower speed [14–17]; this may be an issue if the faulty data has been broadcasted to the rest of the chip, or if the data processing flow does not allow such interruption.

To cope with this issue, it has been proposed in [19,20,21] to anticipate the system failure by checking the monitored flip–flop timing slack by inserting a local prevention delay. If these solutions avoid replaying faulty computations, they do not allow detecting fast voltage drops or timing shifts. As a result, reduced timing margins must be considered. The implementation given in [19] triples the area of the sequential element while the observation window in [20] highly depends on the frequency and the implementation proposed in the pioneer work [21] is based on gate level implementation of the monitoring system, which is thus area consuming.

One may be aware that this concept does not allow catching fast dynamic variations whose effect would imply delay fluctuations greater than the prevention taken.

Within this context, the contribution of this paper is twofold: a new monitoring structure and its associated design flow. The monitoring system, in line with [15,19,20] concepts, aims at anticipating timing violations, induced by process, temperature and slow voltage drifts, over a wide range of operating conditions. This new timing slack monitor may allow the application of dynamic voltage and/or frequency scaling as well as body-biasing strategies, or, at least, provide valuable information to a global monitoring system integrating voltage, temperature and process sensors.

The proposed system detects locally critical timing slacks and monitors on the fly of their evolution with PVT variations or ageing phenomena such as NBTI or HCI [7]. One key feature of the system is the generation of the detection window. Indeed, contrarily to [20] that uses the falling edge of the clock, the detection window is generated by specific Clock-tree Cells (*CC*) directly integrated into the clock network. This solution allows fine tuning the position in time and the width of the detection window. Note however that the insertion of these CC must be done with specific care to avoid tedious iterations in the design flow and not to deteriorate the clock skew.

The rest of this paper is organized as follows. Section 2 gives an overview of the whole monitoring system. The two specific standard cells (the sensor cell and the Clock-tree cell) required to integrate such a monitoring system are then described in detail in Section 3. Section 4 introduces the integration flow that has been used to integrate the proposed monitoring system, including the CC insertion into the clock-tree. Section 5 gives some validation results related to the integration of the monitoring system in a 32 bits VLIW DSP designed in a 45 nm technology. A conclusion is finally drawn in Section 6.

2. Monitoring system concept

Fig. 1 shows the proposed monitoring system, which is composed of two standard cell library elements: a sensor and a specific Clock-tree Cell (*CC*). The sensor is intended to be inserted close to a D-type flip–flops (*DFF*) located at the endpoints of critical timing paths of the design, while *CC* are inserted at the associated clock leaves.

The sensor, acting as a stability checker, is directly connected to a datapath output, i.e. to a DFF input. It also receives a pulse *CP* defining the observation window *DW*, of duration *dw*, provided periodically by *CC*. The edges of *CP* are in phase with *CLK_DFF*, the flip–flop clock also generated within *CC*. The basic function of the sensor is to detect the occurrence of a full or partial transition of the signal *In_A* during this detection window. When this event occurs, the monitor latches a warning signal (i.e. *QN* switches from V_{dd} to 0).

To get the effective detection window DW_{eff}, of width dw_{eff}, internal propagation delays of the sensor have to be considered (Fig. 2).

In_A-to-CP_r and *In_A-to-CP_f* are the sensor internal delay differences between internal edges produced by *In_A* and *CP* rising and falling edges, respectively. *CP-to-CLK_DFF* is the delay between *CP* and *CLK_DFF* rising edges. With such notations, the effective detection window starts at [*In_A-to-CP_r*+*CP_r-to-CLK_DFF*] before the rising edge of *CLK_DFF* node and ends at [*In_A-to-CP_f*−*CP_f-to-CLK_DFF*] before the falling edge of *CLK_DFF*. Thus the effective width dw_{eff} of the effective detection window DW_{eff} is fixed by both the internal structure of the sensor and the timing characteristics of the CC element. More precisely the effective width $dw_{eff} \approx (CP\_r\text{-}to\text{-}CLK\_DFF + CP\_f\text{-}to\text{-}CLK\_DFF) + (In\_A\text{-}to\text{-}CP\_r - In\_A\text{-}to\text{-}CP\_f) = dw + (In\_A\text{-}to\text{-}CP\_f - In\_A\text{-}to\text{-}CP\_f)$. Note that at first order $dw_{eff} \approx dw$.

A key point here is that [*In_A-to-CP_f*−*CP_f-to-CLK_DFF*] must be slightly greater than the setup-time *Tsetup* of the monitored DFF, or at least equal, if one expects detecting timing warnings rather than timing errors.

To take into account the uncertainties on the setup-time (*Tsetup*) estimations (obtained during the design steps), a design guard margin *Gm*, on the value of *Tsetup* can be added. In this case, if during the last clock cycle, the timing slack *Tm* (before occurrence of a setup-time violation) is lower than [*In_A-to-CP_r*+*CP_r-to-CLK_DFF*−*Tsetup*−*Gm*] and greater than [*In_A-to-CP_f*−*CP_f-to-CLK_DFF*−*Tsetup*−*Gm*], the

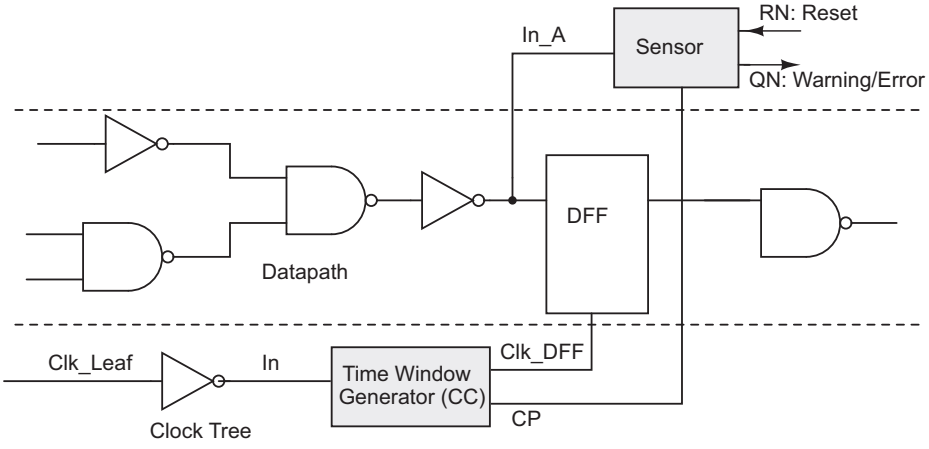

Fig. 1. Monitor system implemented on a path.

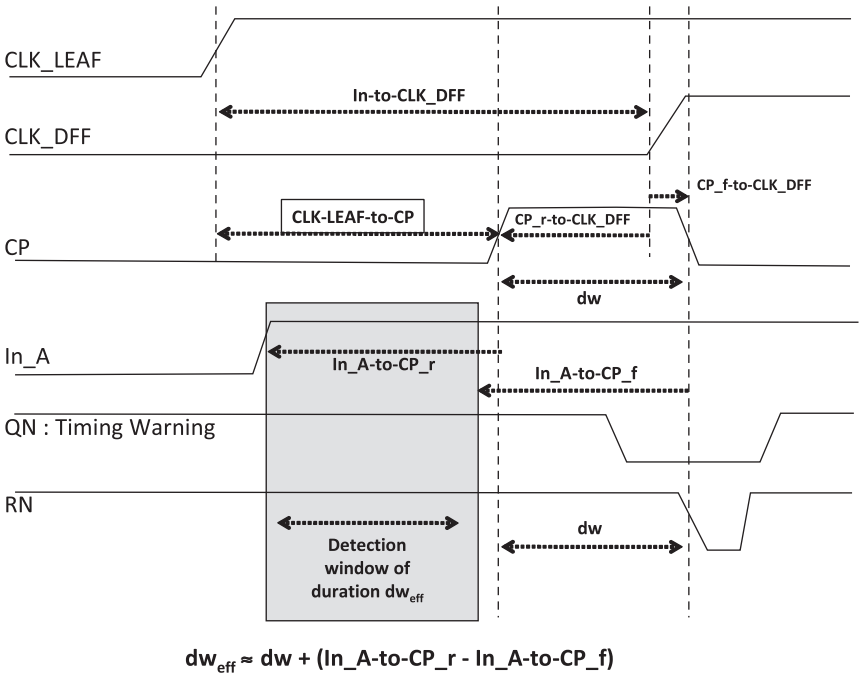

Fig. 2. Transition detection chronogram: Node1_2_Node2 denotes the delay between transitions of Node1 and Node2, while '_r' and '_f' denote rising and falling edges, respectively.

sensor will warn the system that it is close to a timing failure by detecting the occurrence of a signal transition during the detection window.

It is important to keep *Gm* low. Indeed, fixing *Gm* to a positive and significant value results in introducing a small time window, called 'silent window' between the effective detection window and the setup window. In that case, if *Gm* is too large, a fast and undetected transition may occur within this 'silent window'.

On the other hand adopting a negative *Gm* value (to avoid introducing a 'silent window') may lead the system to only warn a timing error without any previous warning. Indeed, if $|Gm|$ is large, the effective detection window overlaps significantly with the setup-time window. As a result, there is a non-null probability to detect a transition that does not meet the setup-time constraint.

This sensor can also be used to detect the occurrence of timing errors without integrating any clock-tree cell, simply by replacing *CP* by *CLK*, i.e. by reducing [*In_A-to-CP_r+CP_r-to-CLK_DFF*] to *In_A-to-CP_r* time. However, in that case, the detection window is too wide (*dw* is equal to half the clock period) and the issue is the same that in [14–17], buffers have to be inserted to compensate short delay paths ending at the input of the monitored flip–flop in order to avoid false error detection.

Finally, to tackle the complexity of actual SoCs, made of blocks, with different timing constraints, or working under different operating conditions (multiple clock or V_{dd} domains), several clock-tree cells CC, implementing different window widths *dw* and different design guard margins *Gm*, may be available in the library for efficient and practical design.

3. Monitoring system blocks

This section aims at introducing the basic operation and the main characteristics of the two specific standard cells required to integrate efficiently and automatically the proposed monitoring system: the sensor (Fig. 3) and *CC*, the clock-tree cell (Fig. 6).

In this sub-section, the sensor schematic representation and its basic operation are first described. Its performances are then reviewed considering a 45 nm technology.

3.1. Single input sensor design and behavior

A single input sensor is composed of a transition detector and a weak feedback loop latch is represented in Fig. 3. The circuit is based on the charging and discharging of the node C, which is the input of the latch.

As shown in Fig. 3, a transition occurring on the input *In_A* of the sensor is propagated by an inverter chain to T1–T4, such that either T1/T2 transistor stack (rising edge) or T3/T4 transistor stack (falling edge) are turned on during a short time interval.

As a result and as illustrated in Fig. 5, the input transition is detected by the sensor if and only the input transition is such that the aforementioned time interval overlaps sufficiently with *CP* pulse to allow the full discharge of node C. In any other case, the node C remains biased to V_{dd}.

To operate correctly, the proposed sensor must be designed according to a main guideline: transistors must be sized to allow a full discharge of node C within the delay of the three inverter chain, since this delay corresponds to the time interval, during which either T1/T2/T5/T6 or T3/T4/T5/T6 are on simultaneously.

3.2. Area considerations

Considering complex synchronous circuits, a large number of observation points are required to obtain an efficient monitoring system. Reducing the area and power consumption of monitor structures is therefore mandatory to detect or anticipate timing failures at a reasonable cost.

To reduce the number of transistors and thus the area overhead, the architecture of the sensor has been defined such as the latch and the tail of the discharge path can be common to multiple transition detectors (Fig. 4). Table 1 compares the

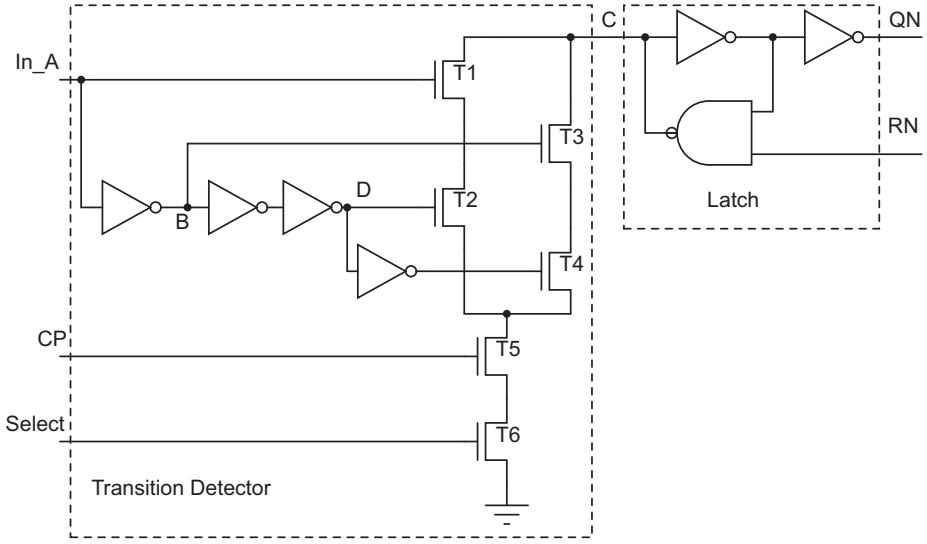

Fig. 3. 1-Input-sensor.

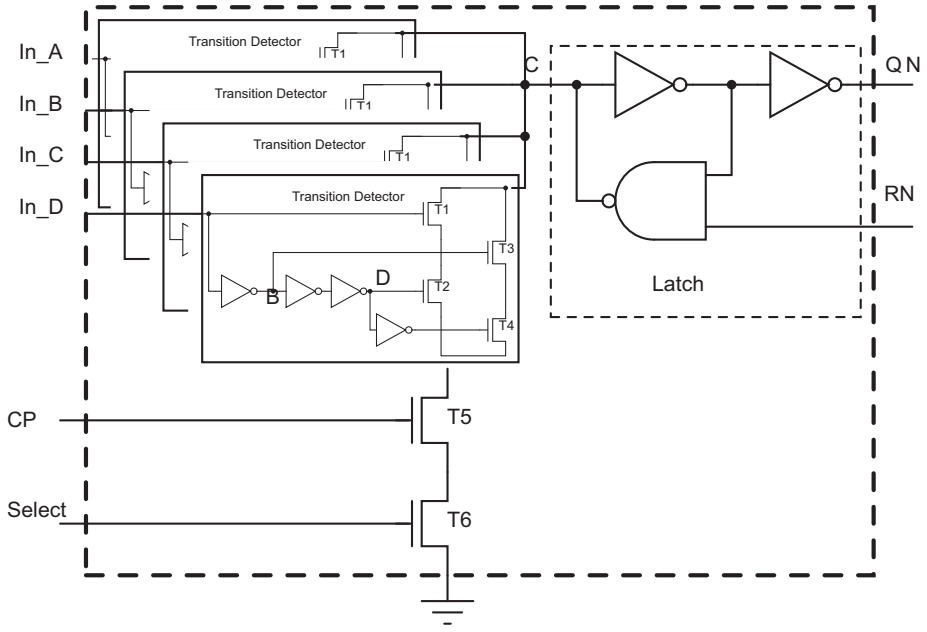

Fig. 4. 4-Input-sensor.

resulting complexity with standard flip–flops and previously proposed solutions. This allows, as indicated in Table 1, to monitor the inputs of one bit with a reduced number of transistors compared to a flip–flop redundancy technique [20].

However, the sharing of the latch has a drawback. It limits the timing performances of the transition detection by increasing the output load of node *C*. Thus, a technology dependent trade-off between timing performance and area (maximum number of inputs) must be found.

Table 1
Additionnal number of transistors per monitored bit.

1-Input-sensor (1 bit)	2-Input sensor (2 bits)	3-Input sensor (3 bits)	4-Input sensor (4 bits)
22	34/2=17	46/3=15.3	58/4=14.5
Min DFF (1 bit)	Min scan DFF (2 bits)	RAZOR2 [15] (1 bit)	[20] (1 bit)
26	34	37 (14 For the latch)	28

3.3. Timing performances

Several timing metrics are used to characterize the behavior of the proposed sensor architecture and thus to ease its integration in actual design flows.

In_A-to-CP time is a first specific timing characteristic of the proposed structure that has similarities with the setup-time of DFF. As shown in Fig. 5, the time spent by the rising transition of *In_A* to propagate through the inverter chain fixes a time window in which T1/T2 (T3/T4 in case of a falling transition) are on simultaneously. Three cases must then be considered, assuming the sensor active i.e. T6 on.

The case (a) corresponds to a situation in which the time interval separating the risings of *In_A* and *CP* allows transistors

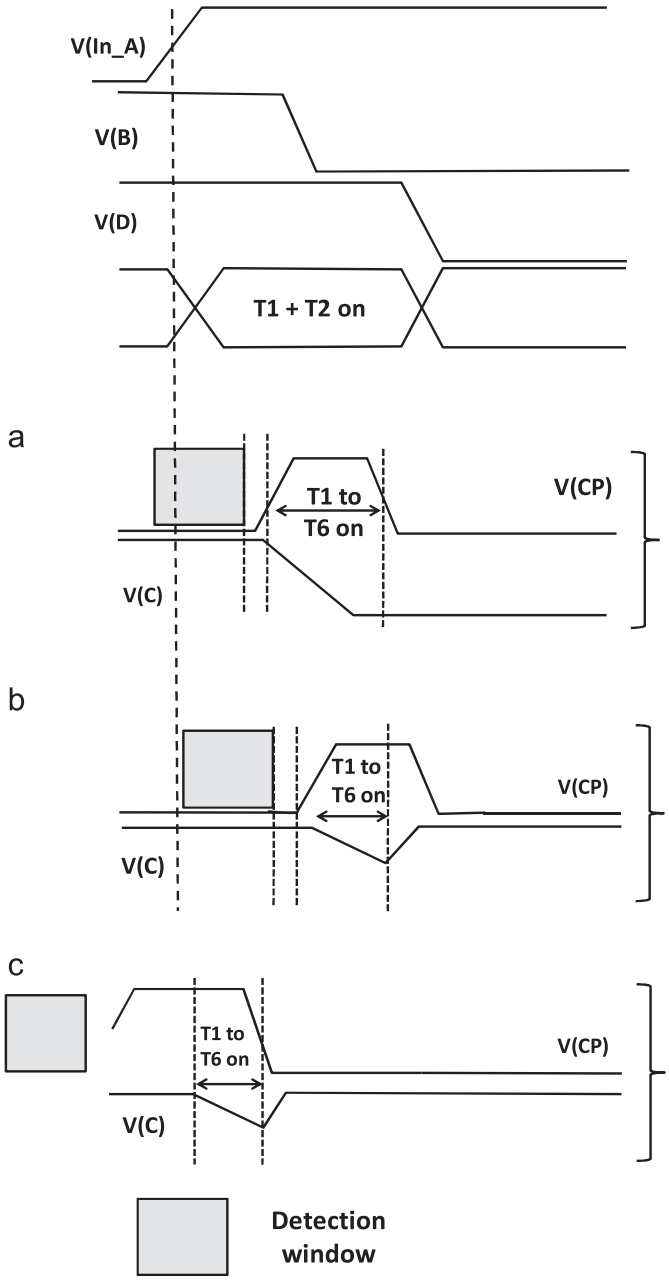

Fig. 5. Chronogram illustrating the effect of *In_A-to-CP* time that fixes the detection window. B and D are internal nodes of the sensor (see Fig. 3).

T1–T5 to be simultaneously on enough time to completely discharge the node *C* and thus detect the input transition.

The case (b) corresponds to a situation in which the time interval separating the risings of *In_A* and *CP* is too long to allow a full discharge of node *C* i.e. the detection of the input transition.

Finally, the case (c) corresponds to a situation in which the time interval separating the rising edges of *In_A* and *CP* is too short, resulting in a too short time window, during which *T1–T5* are simultaneously on. In that case again, the transition is not detected.

Considering these three cases, *In_A-to-CP* timing characteristic defines the maximum time interval that must separate the rising edge of *CP* and the input transition *In_A* to obtain a correct operation of the transition detector. Thus, *In_A-to-CP* and *CP* pulse duration *dw* times fix the bounds of the effective detection window that, respectively, starts and ends at [*In_A-to-CP_r+CP_r-to-CLK_DFF*] and [*In_A-to-CP_f − CP_f-to-CLK_DFF*] before the rising edge of *CLK_DFF*.

RN-to-QN time is a second specific characteristic of the proposed monitoring structure. It is related to the time required to set the internal node *C* to a logical '1'. It is an important timing characteristic, which must be as low as possible to be compatible with high speed designs.

Similarly to *In_A-to-CP* and *RN-to-QN* times, two other timing characteristics must be satisfied for a correct operation of the proposed structure: the minimum pulse widths of *CP* and *RN*, respectively.

Finally, two last timing metrics, characterize the time spent by the proposed structure to output a timing warning (or an error) depending on the signal arrival order. If *CP* arrives first then, *CP-to-QN* characterizes the speed of the monitor, while *In_A-to-QN* is a better indicator if the input transition occurs first.

Table 2 gives simulated timing performances (in ps) of the proposed monitor structures. Note that these performances are reported in cases of 1, 2, 3 and 4 transition detectors sharing the same output latch. They have been obtained considering the worst case process conditions, a supply voltage of 1.05 V, a temperature of 125 °C, an output load (net *QN*) of 10 fF and an input ramp duration of 100 ps (*In_A*) for both falling and rising transitions.

Note also that the simulated structures were designed to optimize the *In_A-to-QN* timing characteristic, while maintaining the area as small as possible.

As shown in Table 2, increasing the number of transition detectors and sharing the same output latch results in an increase of *RN-to-QN*, *CP-to-QN* and *In_A-to-QN* times. This timing degradation (mainly due to the increase of the capacitance of node *C* and other internal nodes) remains reasonable up to four transition detectors sharing the same output latch.

Table 2 also discloses a difference of roughly 50 ps between the rising and falling edges for the *In_A-to-CP* timing characteristic at almost all *V,T* operating conditions. However, with a *CP* pulse duration of at least 150 ps, this timing asymmetry does not cause functional issues for the system.

Table 3 reports the evolution of *In_A-to-QN* (for which the 4-input sensors operate correctly) w.r.t. the different *V*, *T* conditions in the worst case process. These results predict a correct operation with a supply voltage as low as 0.6 V over the whole

Table 2
Timing characteristics (ps).

	# MOS	*In_A-to-QN* rise/fall	*RN-to-QN* fall to rise	*CPmin* high	*RNmin* low	*In_A-to-CP* rise/fall
S1	22	156/203	221	165	165	63/98
S2	39	263/302	404	163	215	53/102
S3	46	278/322	531	167	280	54/110
S4	58	295/359	573	164	320	48/105
Min DFF	26	225/230	X	210	X	X
Min scan DFF	34	235/242	X	300	X	X

Table 3
In_A-to-QN time (ps) w.r.t. *V*, *T* conditions of a 4-input sensor.

V/T	−40 °C	10 °C	60 °C	85 °C	125 °C
0.6 V	4002	2785	2464	2195	1819
0.8 V	741	710	672	646	602
1.0 V	360	333	329	326	320
1.1 V	274	257	257	258	257
1.2 V	224	212	214	215	216

temperature range [−40 °C; 125 °C]. This illustrates the robustness of the sensor against PVT variations. Note that the 1- to 3-input sensors exhibit better performances.

3.4. Power consumption

Worst power consumption estimations are typically obtained considering the best process corner, and high temperature (125 °C) and supply voltage (1.2 V) values. According to these conditions, it was found that the worst static power consumption, obtained considering all possible biasing conditions and 1 to 4-input sensors is about 100 nW (1 nW for the typical case). These values are similar to the one of a minimum drive DFF.

The dynamic power consumption of the proposed sensors is mainly dissipated during the switching of the internal inverter chain of the sensor (Fig. 3). Indeed, these inverters have the same activity ratio than the monitored data, while the node *C* and the latch switch only when a transition occurs within the time detection window, i.e. when a timing warning occurs. Table 4 lists values of the dynamic power consumption for 1-input to 4-input sensors during detection of a transition.

Table 4
Dynamic consumption characteristics (nW/MHz).

S1	S2	S3	S4	Min DFF	Min scan DFF
7.6	15	22	30	8	9

As shown, the dynamic power consumption of the proposed sensors increases linearly with the number of inputs. This expected result is due to the unshared inverter chains that consumed the main part of the overall dissipated energy. Thus, regarding dynamic power consumption, adopting a multi-input structure is not very effective.

The dynamic power consumption dissipated during a reset of the latch is due to the slow charging of node *C* by a small PMOS transistor. This consumption also increases linearly with the number of inputs, due to the linear increase of *C* capacitance value. In worst consumption case, with a *RN* transition time of 160 ps, the power consumption is between 9.2 nW/MHz for a 1-input-sensor and 14 nW/MHz for a 4-input sensor. However, as resetting the system is not a frequent task, this consumption has a minor impact on the overall performance of the proposed monitoring system.

3.4.1. Clock-tree cell design and characteristics

This paragraph aims at introducing the proposed clock-tree cell of Fig. 6 and its associated characteristics in a 45 nm technology.

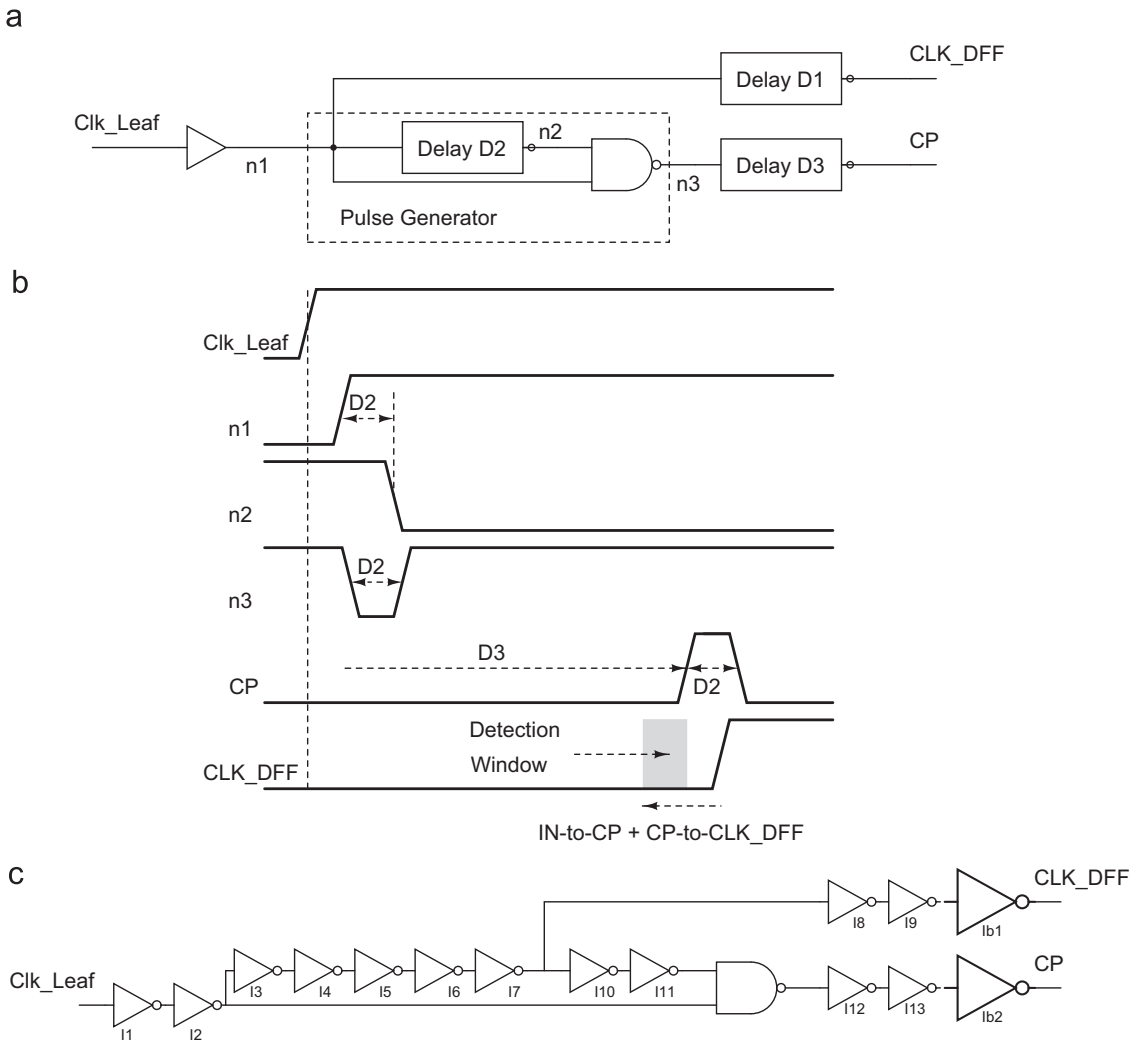

Fig. 6. Clock-tree cell *CC* implementation and its associated chronogram.

3.5. Clock cell design

The generation of *CP* pulse of duration *dw* is a key design step, since it fixes in time the effective detection window DW_{eff} and defines its duration dw_{eff}. Thus, specific *CC* clock cells have been developed.

As shown in Fig. 6a, these cells have to be inserted at the leaves of the clock-tree and more precisely between *CLK_LEAF* and *CLK_DFF* that correspond, respectively, to a clock leaf and the clock input of the monitored *DFF*.

Fig. 6a gives the basic structure of *CC*. It is composed of a pulse generator and delay elements. As shown in Fig. 6b, the pulse generator fixes the *CP* pulse width i.e. the duration *dw* (equal to D2 delay) of the detection window that may be reduced if required to the delay of a single inverter.

Delay elements *D*1 and *D*3 fix the time interval *CP-to-CLK_DFF* separating the rising edge of *CLK_DFF* from the rising edge of *CP*, i.e. the position in time of the detection window w.r.t. the rising edge of the clock that starts at *In-to-CP+CP-to-CLK_DFF* (see Fig. 2).

Since *CC* cells fix the position in time and the width of the detection window, their sensitivity to PVT conditions has to be as low as possible. This is done by using as far as possible identical elements on the different electrical paths, and sharing as many elements as possible between the two different paths (*CLK* and *CP*). This solution is illustrated in Fig. 6c Inverters *I*1 to *I*7 are shared by *CLK_LEAF* to *CLK_DFF* paths and *CLK_LEAF* to *CP* path, while *I*8, *I*9 and *Ib*1 are identical to *I*12, *I*13 and *Ib*2. Thus most of the sensitivity difference between those paths is due to *I*10, *I*11 and the two-input NAND gate. Some results related to PVT sensitivities of the *CC* are given in next paragraphs.

3.6. Area considerations

To further reduce the area overhead of the monitoring system, *CC* clock cells may also be shared by different single input or multiple input sensors as shown in Fig. 7. However, in such a case, the clock-tree synthesis has to cross the *CC* cells and to perfectly balance the clock and *CP* skew, particularly between the instrumented flip–flops.

3.7. Timing performances and behavior

As for the sensor, several timing metrics are used to characterize the behavior of *CC* and ease the flow integration.

Among them, *CLK_LEAF-to-CLK-DFF* and *CLK_LEAF-to-CP* are timing metrics characterizing the propagation delays of the incoming clock edge to *CLK_DFF* and *CP* outputs, respectively. These timing characteristics, especially *CLK_LEAF-to-CLK-DFF*, must be characterized in order to efficiently control the clock-tree timing skew and more precisely to be able to reduce the skew between standard clock leaves and leaves ended by a *CC* clock cell.

The time interval between edges of *CP* and *CLK_DFF* (*CP-to-CLK_DFF*) and *CP* pulse width *dw* are also key timing metrics of *CC* clock cells. Having several clock cells with different programmable timing characteristics, in a standard cell library will ease the whole monitoring system integration.

Table 5 gives the post layout simulated mean and standard deviation values of timing characteristics of a *CC* cell with high driving capabilities. In this design, thanks to programmable delays, the pulse width *dw* of *CP* may be set to two different values, dw_1 and dw_2, corresponding to two effective detection windows DW_{eff1} and DW_{eff2}, respectively. These Monte Carlo simulations values have been obtained considering a supply voltage of 1.1 V, a typical process, a temperature of 25 °C, a

Table 5
Timings of a programmable CC (ps).

	DW_{eff1} mean/std dev	DW_{eff2} mean/std dev
CLK-LEAF-to CLK_DFF	231/17	231/17
CLK-LEAF-to-CP	224/16	173/13
CP(Rise)-to-CLK_DFF	8/5.2	61/8
CP(fall)_to_CLK_DFF	−98/−10	−153/−13
dw	106/11.3	214/15.3

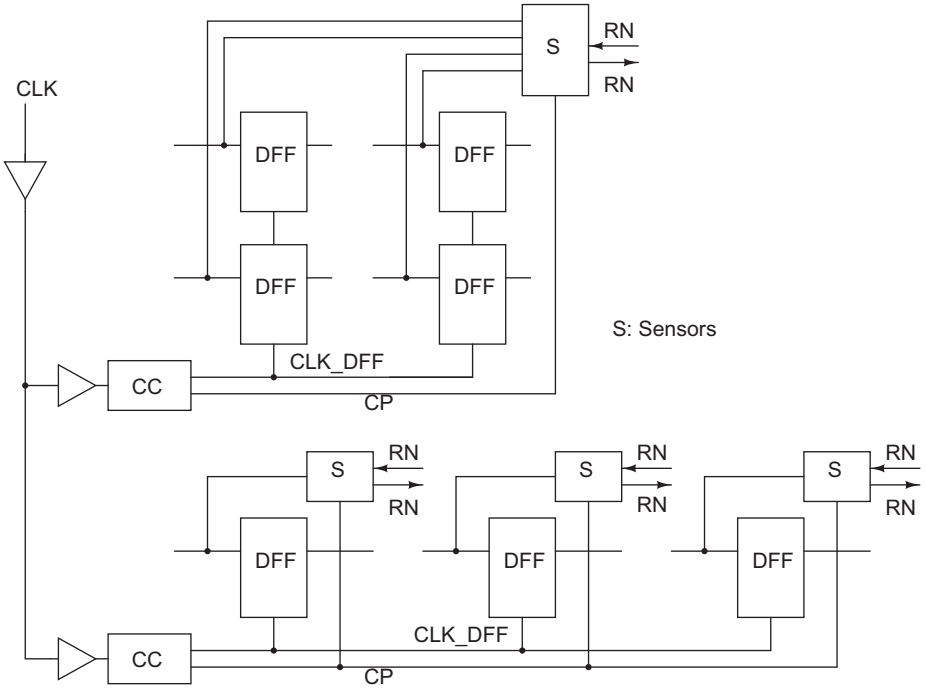

Fig. 7. Distributed clock-tree and the proposed monitor system.

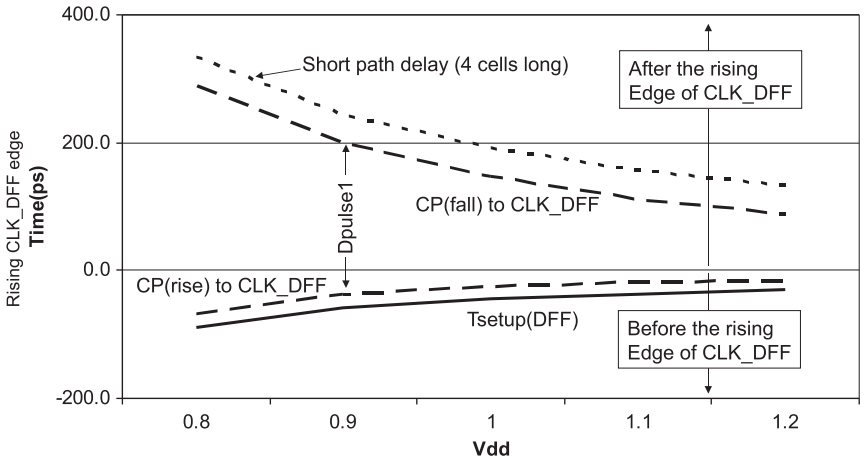

Fig. 8. Simulated *CP* edges arrival times (w.r.t *CLK_DFF* rising edges) for DW_{eff1} time window generated by the programmable *CC* (typical process corner at 25 °C).

CLK_LEAF transition time of 30 ps, output loads of 60 fF for *CLK-DFF* and 30 fF for *CP* and with 1000 runs.

As shown in Table 5, dw_1 *and* dw_2 are equal to 106 and 214 ps, respectively. This allows detecting transitions on a short time interval representing 5% of the period of a 500 MHz clock signal. In the same way, one may notice that *CP(Rise)-to-CLK_DFF* ranges 8–61 ps in case of a rising edge of *CP*.

Fig. 8 represents the evolution with V_{dd} of the rising and falling edges arrival times of *CP* (with respect to the rising edge of *CLK_DFF*) and thus the evolution of *dw*1 related to the detection windows DW_{eff1}. Note that this effective detection is positioned in time *In_A_to_CP* ps before the rising edge of *CP*.

Firstly one may observe that the lower the supply voltage value , the wider the *CP* pulse width and *CP-to-CLK_DFF* time are in an absolute value. However, dw_1 sensitivity to V_{dd} is reduced ($<$ 55 ps/100 mV). It is quite similar (parallel) to the one of a short delay path (made up to 4 or 5 serially connected cells) and is significantly lower than the one of high logic depth paths (10–20 cascaded cells). This behavior is a key advantage to maintain a detection width relatively small compared to the propagation delays of critical paths.

It can also be observed that *CP_r_to-CLK_DFF* sensitivity to V_{dd} is parallel to the one of a standard DFF setup-time (*Tsetup*) meaning that the position (w.r.t. to the *CLK_DFF* rising edge) of the effective detection window DW_{eff1} evolves accordingly to the setup-time value. This is a key point to warrant that the system monitors, over its complete operating range, timing warnings and not timing errors.

3.8. Power consumption aspects

Since *CC* cells are placed at some leaves of the clock-tree, their activity factor is identical to the one of any clock-tree buffer. Their power consumption must therefore be kept as low as possible. Table 6 gives figures of merit of (i) the designed clock cells with large driving capabilities and (ii) some buffers (a small drive and a large drive) dedicated to the design of the clock-tree. The reported values have been obtained considering the best process corner, for a temperature and a supply voltage of 125 °C and 1.2 V, a *CLK_LEAF* transition time of 30 ps, output loads of 60 fF for *CLK-DFF* and buffers outputs, and 30 fF for *CP*.

From Table 6, one may conclude that *CC* dissipates, at least 8 times more leakage power than a large drive clock buffer, and roughly 15 times more dynamic power consumption than a large drive buffer. However, one must keep in mind that the number of *CC* will remain significantly lower than the number of clock-tree buffers in a design.

Table 6
Power consumption of CC.

	Static power (mW)	Dynamic consumption (µW/MHz)
Small drive buffer	8.8×10^{-5}	5×10^{-3}
Large drive buffer	5.4×10^{-4}	25×10^{-3}
Clock cell 1 (1 detection window)	2.0×10^{-3}	38×10^{-2}
Clock cell 2 (2 detection windows)	2.2×10^{-3}	40×10^{-2}

3.8.1. Discussion

In this section, the schematic representation, behavior and performances of the sensors and associated clock-tree cell, required to integrate the proposed monitoring system, have been analyzed considering a 45 nm technology.

This analysis has demonstrated the ability of the sensor, designed in 45 nm technology, in detecting, within a time window as short as 160 ps (see *CPmin* in Table 2) a transition occurring on its input, on a wide range of process, temperature and supply voltage conditions.

Similarly, the results obtained have demonstrated that the proposed *CC* is an interesting structure to deliver to the sensor suitable *CP* pulses, since its sensitivity to the supply voltage is quite similar to that of a short path, and to that of the setup-time of DFF available in the considered 45 nm technology (see Fig. 8).

At this point, one question remains: how is it possible to integrate automatically and adequately these cells, and thus the proposed monitoring system in complete circuit? Next section addresses this question.

4. Monitoring system integration flow

To demonstrate the whole system efficiency at the circuit level, both in terms of performance and integration easiness, a test chip has been designed accordingly to a dedicated design flow. This section aims at describing the specific integration flow adopted to validate the monitoring system on a real test case.

4.1. Design flow overview

The integration flow adopted to validate the timing slack monitoring system is shown in Fig. 9. Three steps are necessary to integrate it efficiently.

In a first step, a prototype (a placed and routed design) is obtained. It allows identifying the critical paths to be monitored, and to get a first description of the clock-tree. Sensors are then inserted in the netlist. Then, starting from the prototype obtained from step 1, final place and route steps are performed, considering additional timing constraints related to *CC* and sensors insertion. In this last step, balancing the nets between monitored DFF and their associated sensors is a key design guideline. Step 1 corresponds to a conventional integration flow and is thus not detailed afterward.

4.2. Critical path choice

Getting an exhaustive timing monitoring of a chip is unrealistic. Indeed, this would result, at first order, in doubling the die area occupied by sequential elements and in a dramatic increase of the power consumption. Thus, the most relevant critical paths to be monitored have to be selected.

One possible solution is to use Statistical Static Timing Analysis [1,3]. Indeed, an SSTA is performed to identify the paths (assuming that these paths are correlated enough) with the highest probabilities of violating the setup-time constraint. However, with the decreasing transistor length and the increasing impact of local and random variations, correlations between timing paths shrink down, leading to a growing number of paths that have to be monitored due to their different timing behaviors.

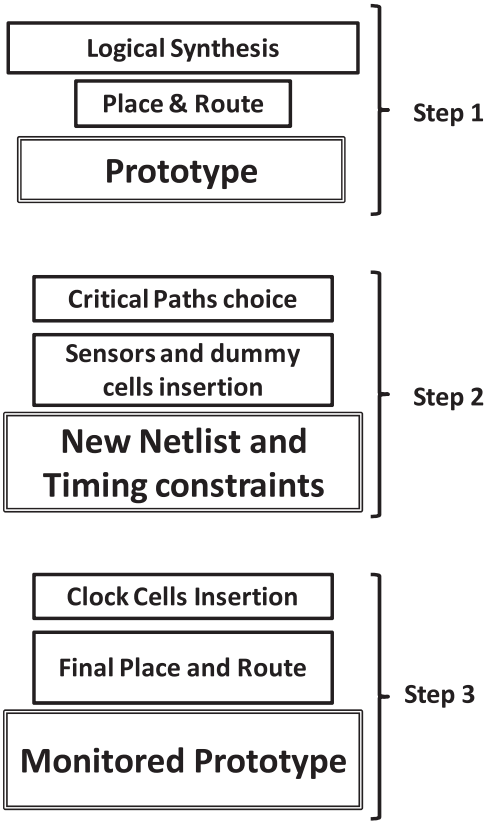

Fig. 9. Flow to integrate the proposed monitoring system.

Therefore, another selection policy was adopted since our goal was to prevent a system failure by anticipating the occurrence of any timing violation. Considering this constraint, our policy was to impose specific target timing slack constraints during the synthesis and the place and route steps.

More precisely specific target slack constraints were chosen to create a reduced set of pre-defined critical paths to be monitored as shown in Fig. 10. The target slack constraint of this set of critical paths was relaxed with respect to 'slack value 2' rather than 'slack value 1'. This set of paths was chosen such as to include the worst critical paths identified during several synthesis runs performed considering the worst timing corner and tight timing constraints. Although we have not used it, an SSTA might also bring useful complementary information for this selection.

This results in a timing slack distribution characterized by two distinguishable sets of paths (Fig. 10): a reduced set of paths, characterized by reduced timing slack values (slack value 2), forms the set of paths to be monitored, and a huge set of paths characterized by larger timing slack values (slack value 1), and more precisely by paths with a slack greater than roughly 3 times the standard deviation of the worst path delays of the considered design. This criterion has been adopted in order to warrant that a non-monitored path will never have a greater delay than a monitored one due to some process variations even if the statistical correlation between the paths is reduced to 0.

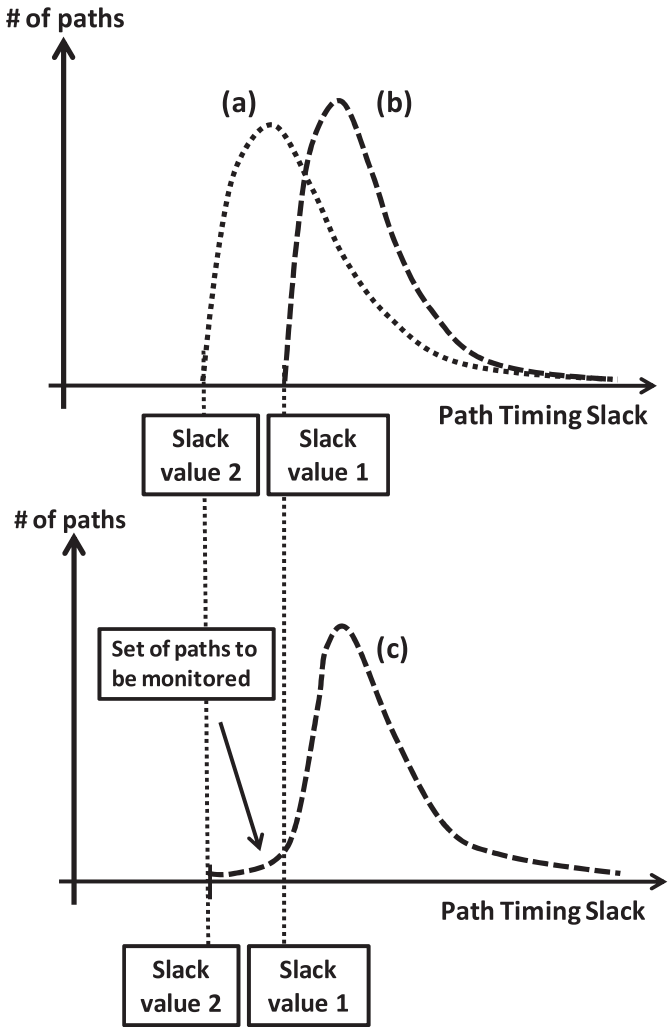

Fig. 10. (a) Timing slack distribution obtained for target slack value 2, (b) for target slack value 1 and (c) after definition of the set of paths to be monitored.

An illustration is given in Table 7. The results are related to the application of the proposed flow to a real design, a low power 32 bits VLIW reconfigurable DSP (considered in the next section for the validation). In applying these constraints during a synthesis flow, we can drastically decrease the amount of endpoints to observe: from 3695 to only 24 on a 100 ps interval.

Such timing slack distributions may be obtained, thanks to multi-mode capabilities of tools and/or to specific relaxing tool commands. However, it is of prime importance to warrant that the paths to be monitored are representative of the circuit in any possible operating conditions. Thus, a verification step is required in order to fulfill this condition, using either multi-corner or statistical simulations.

This specific flow is also equivalent to over-constraining some paths from a timing point of view, if the original placed and routed prototype was obtained considering a target slack equal to 'slack value 2'. In the case demonstrated in this paper, the choice has been to keep equivalent worst case operating frequencies, which results in increasing the power consumption as shown in Table 7. This overhead is expected to be significantly lower than gains achieved by applying adaptive voltage, body-biasing and frequency scaling techniques. Note that 'slack value 1' and 'slack value 2' are design parameters allowing other choices, like keeping same power and frequency in typical conditions. Further results will be given in the next application parts.

Choosing the set of paths to be monitored according to their timing criticality is instinctive to capture the overall timing behavior of the circuit. However, this criterion may be insufficient if the selected critical paths have a low activity rate. Indeed, in such a case, the adaptive system (at the decision level) may have to wait for several hundreds of clock cycles to be sure to have an effective and correct monitoring of the timing slack. Thus designers have to verify the paths that are to be monitored have a high enough activity rate, or even choose the paths to be relaxed from a timing point of view according to both timing criticality and activity rate criteria. Note that the definition of the best choice policy is out of the scope of this paper.

Table 7
Number of critical path endpoints and performances penalties (worst timing corner at 1.05 V, 125 °C, logical synthesis results).

	# Critical paths in 100 ps from the WNS	Power (mW)/Maximum Frequency (ns)
Typical flow	3695	29/1.6
New flow	24	35/1.7

4.3. Specific monitoring system cells insertion

This sub-section describes the most important design steps (and their associated timing constraints) allowing an efficient and automatic insertion of the sensors and the clock cell within complex designs.

4.3.1. Sensor cells insertion

The monitoring phase (step 2) has to be performed after an identification of the critical paths, in all operating conditions, but before the final place and route step. As shown in Fig. 9, the monitoring phase aims at providing a new netlist, in which sensors cells have been inserted as illustrated in Fig. 11. This step also provides a new timing constraints file to guide the final place and route step. Note that this can be fully automated by developing specific scripts. Note also that four DFF sharing the same sensor has been found as a practical limit in our test case.

4.3.2. Clock cell insertion

One point that must be noticed is that *CC* cells cannot be inserted during the second step. Their insertion must be done during step 3, i.e. when sufficient data about the final cell placement and about the clock-tree structure are available.

This typically avoids bad placements of the *CC* w.r.t. the DFF and sensors. Such bad placement results in positioning inadequately w.r.t. to rising clock edges, the detection window and in modifying locally the clock skew and thus generating timing violations.

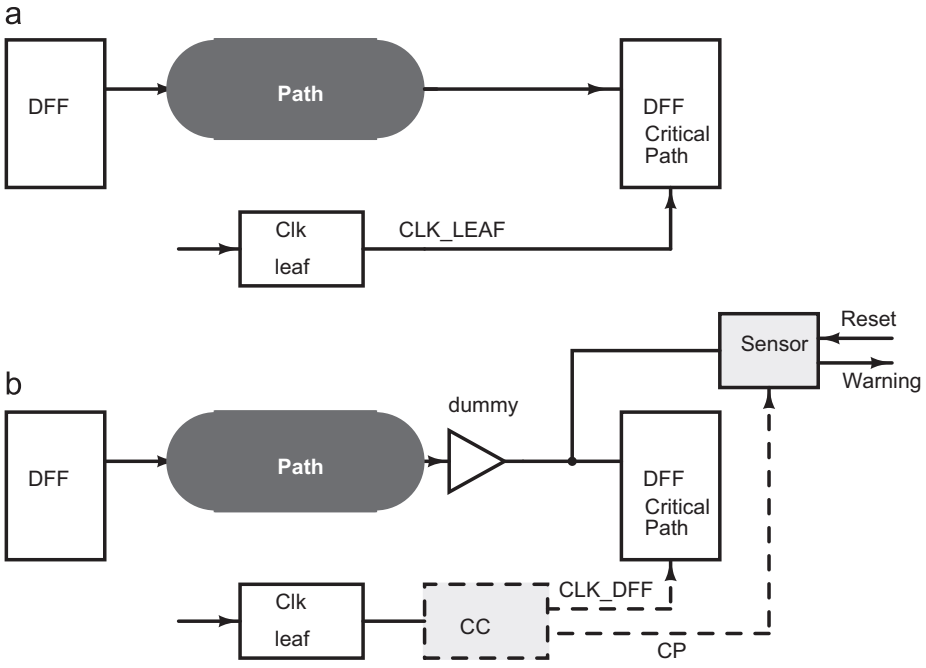

Fig. 11. Critical path before (a) and after (b) the monitoring step. Sensors are inserted with specific dummy cells allowing easing the place and route and satisfying constraints to balance the wires. Clock cell is inserted after, during the place and route step.

CC cell insertion is therefore done, during the final place and route steps, after a first Clock-Tree Synthesis (*CTS*) is performed according to the additional timing constraints defined at step 2. *CC* are inserted at the adequate clock leaves, thanks to the graph interface of the *CAD* tool and dedicated scripts. The insertion done, a second *CTS* is run to further balance the tree, and minimize the global skew of the design, after an update of the timing constraints.

4.3.3. Timing constraints

To avoid the introduction of timing violations, while inserting clock cells during step 3, specific timing constraints must be defined during step 2 (and added to the timing constraints file) and applied during the final place and route steps. Three types of additional timing constraints are considered during the final steps of the proposed design flow.

The first category is made of timing constraints aiming at guiding the placement so that sensors and monitored flip–flops are placed close together. Practically, this is achieved by inserting specific dummy cells (step 2), with a 0 ps delay, at the end of the datapath (Fig. 11). This allows integrating in the timing constraint file, a target delay of 0 ps between the output of the dummy cell and the inputs of the associated DFF and sensor. As a result, the place and route tool optimizes the final placement and routing such as the timing edges, between the dummy cell and the inputs of the related sensor and DFF, have the smallest values as possible.

The second category of timing constraints is related to the Clock-Tree Synthesis (*CTS*). Timing constraints are added in order to guide the first final clock-tree synthesis. After the insertion of *CC* cells, timing constraints related to the clock-tree synthesis are updated to take into account these modifications of clock-tree during the second *CTS*. The two main goals are to be able to achieve the same target clock skew than during the first *CTS*, and also to ensure that *CC* would not be moved in the tree during the second *CTS*.

The third category of timing constraints aims at avoiding the detection of transitions produced by short paths ending at the same points as those of monitored critical paths. This can be achieved by imposing a minimum target delay between the input of those short paths and the output of the path to be monitored. Practically, short paths sharing the same output paths than those of critical ones are constrained such as their minimum arrival time is equal to the ending of the detection window. This is achieved by inserting additional cells along these short paths. However, the reduced values of *CC* delays (compared to [14–17]) and the reduced detection window width enable inserting a limited number of buffers along short paths.

As a result, the three categories of timing constraints allow reducing (by keeping the *CC*, sensors and DFF close together) the impact of spatial process variations on the timings of the different monitoring system elements, and the differential impacts of local voltage drops or temperature gradients on the sensor and *CC* cell timings.

4.4. Discussion

In this section, we have proposed a design flow allowing integrating the proposed monitoring system in complex circuits. Note however that alternative flows may be defined according to the self-tuning [20] or adaptive strategy considered by designers.

This integration flow has been developed to demonstrate that it is possible to integrate the proposed monitoring system within complex circuits using available *CAD* tools. Finding the best self-tuning or adaptive strategies (and their associated integration flow and design metrics) based on the proposed monitoring system, is out of the scope of this paper. Further results related to the application (impact on power, speed, area and clock-tree) of this flow are given in the next section. We therefore assume that it is possible to define an adequate policy at the system level to collect and interpret data provided by the timing slack sensors and other type of sensors such as temperature, voltage and process sensors.

5. Validation

A low power 32 bits VLIW reconfigurable DSP [18] has been used to validate the monitoring system and its associated design flow. This section aims at introducing the results obtained.

The VLIW DSP incorporates 4 SRAM memories, several register banks and different computing elements such as dividers or Multipliers–Accumulators (MAC). Designed in a 45 nm STMicroelectronics Low Power technology, it contains about 13,400 flip–flops, and has a $606 \times 558\ \mu m^2$ core floorplan implementation (Fig. 13). The timing constraint period is 1.5 ns for nominal process conditions and at 25 °C and 1.1 V.

This IP block, used in a telecom SoC, has been considered as an ideal test case to validate our proposal. Indeed, due to the complexity of the considered real-time application, no replay of instruction sets was feasible at the system level, and timing performances have to be high enough to satisfy a strained real-time environment.

6. Path level validation

In a first validation step, we consider the worst critical path, of logic depth equal to 14, to validate the functionality of the monitoring system, and to estimate its robustness to PVT variations. This critical path is one of the original place and route prototype and not the one extracted after an application of the proposed integration flow.

The path topology has first been extracted. Then a sensor and a programmable *CC* have been manually inserted, respectively, at the input of the endpoint DFF and at the clock leaf. A characteristic of the inserted CC was its ability to deliver, on command, two different *CP* pulse widths dw_1 and dw_2 with two different *CP-to-CLK_DFF* times.

The arrival times of the data and both edges of *CP* pulses (for dw_1 and dw_2) as well as the setup-time of the endpoint DFF were extracted from electrical simulations performed at various PVT conditions and, especially at different voltage values. From these simulations, the maximal operating frequency (*Fmax*) of the path was extracted as well as the maximal operating frequency (*Fmax_Sensor*(DW_{effi})) obtained considering outputs of the sensor (the highest frequency value at which no warning was observed). *Fmax_Sensor* to *Fmax* ratio for typical PVT conditions (typical process/1.1 V/25 °C) are, respectively, 94% and 90% for effective detection window DW_{eff1} and DW_{eff2} highlighting the high time resolution of the proposed system. Identical results were obtained at different PVT conditions leading to an averaged *Fmax_Sensor*/*Fmax* value of 96% for DW_{eff1}, demonstrating the correct operation of the whole monitoring system.

In a second validation step, statistical simulations were done to evaluate the probability density function of having a correct behavior of the system, in presence of global and local process variations simulated with a statistical BSIM4 model card. Fig. 12a and b illustrates the results obtained at voltage values of 1.1 and 0.8 V. In these figures, the black lines represent the cumulative density function (*CDF*) of the setup-time with respect to the

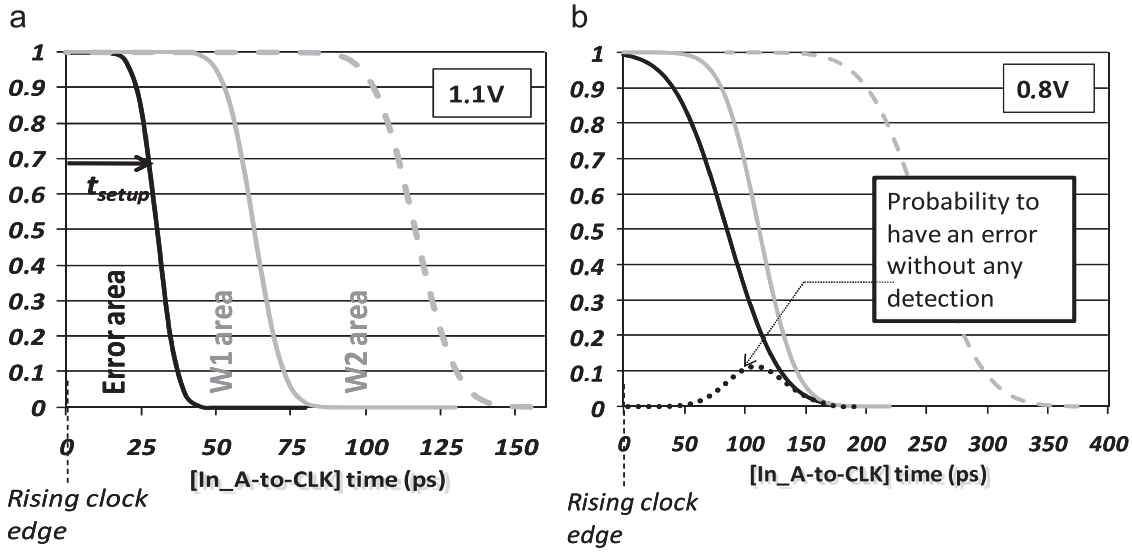

Fig. 12. Proposed monitoring system efficiency w.r.t. the supply voltage value.

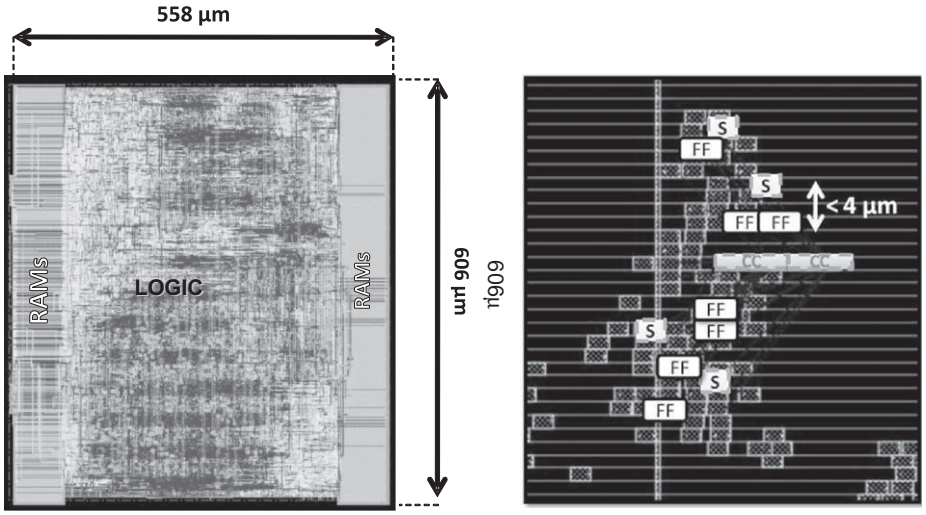

Fig. 13. Final layout and monitor cells location.

arrival time of the rising clock edge at $t=0$ along the x axis. Similarly, the gray and dashed curves are the CDF of the detected incoming transitions within the detection windows DW_{eff1} and DW_{eff2}, respectively. More precisely, these curves give the cumulated probability of detecting a transition arrived at a given abscissa x_0. Thus, if the setup-time CDF does not overlap, along the x-abscissa, with the gray and dashed CDF, the probability that the sensor flags a warning rather than a timing error is 0. This is the case of Fig. 12a. Indeed, at 1.1 V, the impact of process variations is rather small so that the CDF curves are steep, and one can be sure that any transition, detected by the sensor in the time window DW_{eff1}, satisfies the setup-time constraint. On the contrary, at $V_{dd}=0.8$ V (Fig. 12b), the relative variances of the monitoring system timing metrics are important resulting in smoother CDF curves. As a result, the design guard marging Gm adopted is not sufficient to warrant that CDFs do not overlap. Thus, there is a non-null probability to have a detected transition within DW_{eff1} that does not satisfy the setup-time constraint. One must therefore conclude that there is a supply voltage lower bound below, which the system does not operate correctly according to an expected level of confidence. This is not the case for DW_{eff2}. Consequently this result justifies the use of programmable CC to enlarge the voltage range within, which the system may operate and thus to fully take advantage of voltage scaling strategies. Note that this programmability capability may also be used to tune the monitoring system according to the quality of the process or conversely to sort chips according to their speed after fabrication and more precisely during test and characterization steps done at the foundry.

6.1. Circuit level validation

This section describes the main results obtained during the circuit level validation step. These results are related to the proposed integration flow (Fig. 9), and to the efficiency of the monitoring system concept, especially the potential benefits that may be expected from its use in combination with dynamic voltage scaling strategies.

6.1.1. Integration of the monitoring system

Starting from a placed and routed prototype of the considered arithmetic and reconfigurable block, 160 critical path endpoints, defining a set of critical paths, mostly situated at the output of the

first stage of the pipelined MAC, were isolated, by relaxing their timing constraints.

The maximum operating frequency *Fmax* related to these path endpoints, and thus the one of the circuit, was characterized in nominal conditions. Values of 1.78 and 1.36 ns were found as the minimal periods allowing correct timing behaviors of this set of paths in the nominal and best timing corners, respectively.

Aiming at reducing the number of endpoints to be monitored, we ran a multi-corner analysis to refine this initial selection. Finally, to select the reduced set of endpoints to be monitored, we analyzed successively the timing behaviors of the worst, the 10 worst, the 25 worst and finally the 50 worst endpoints deduced from timing analysis in nominal conditions (typical process, 1.1 V, 25 °C). These sets of paths were then compared to the equivalent sets of paths obtained at several PVT conditions.

Table 8 gives results obtained. Note that each box of this table reports three values. The upper one indicated the number of worst endpoints obtained in typical conditions (typical process, 1.1 V, 25 °C) being among the n worst considered endpoints ($n=1$, 10, 25, 50).

As shown, the worst endpoint obtained in typical conditions does not remain the worst one in all conditions. Thus monitoring only this single endpoint is insufficient to cover the timing behavior of the circuit.

Considering the 10 worst endpoints obtained in typical conditions rather than the worst one improve results since at least seven (i.e. 70%) of these endpoints are among the 10 critical endpoints at all other PVT cases. This result can be read in the second line of Table 8. We further increased the number of considered critical endpoints to 25 and finally 50. None of these endpoint subsets are always the most critical subset in all PVT conditions. However, we noticed that the worst endpoint associated to all considered PVT conditions was always in the sets of 10, 25 and 50 worst endpoints obtained in nominal condition. We thus conclude that monitoring these 10, 25 or 50 worst endpoints, rather than 160, was sufficient to warrant monitoring the true worst endpoint in all PVT cases. Considering these results, we decided to monitor the 50 worst endpoints to cover a larger window of arrival times; as shown in Table 8. The greatest negative slack, considering 50 endpoints, is 1.3 to 2.7 times larger than the one obtained considering only 10 paths.

Analyzing the clock-tree architecture is mandatory to know, which CC standard cells have to be used and where they will be inserted. The clock-tree synthesis applied in our specific flow led to a 15 level clock-tree, with 1404 sub trees, monitoring 13,400 flip–flops through 530 clock gating elements. The latency from the beginning of the tree to leaves was 1.19 ns in the nominal corner condition, with a skew of 102 ps. These values were captured at the very end of the prototype place and route.

Considering the endpoints to be monitored, and according to a design strategy aiming to minimize the area overhead (close path endpoints are gathered on a same sensor, with a limit of 4), the number of inserted sensors in our design was 19 (to monitor the 50 endpoints), with 11 *CC*.

The final place and route step (with two clock-tree synthesis and new timing constraints) led to a final clock skew of 158 ps and a latency of 1.28 ns in nominal conditions (Table 9). Note that results reported in Table 9, as well as all results given afterward have been obtained with SoC Encounter Power and Timing integrated tools.

Analyzing the final placement of DFF and sensors, we found that the maximum distance separating DFF from its sensor was 4 μm (i.e. about 2 standard cell height) as shown in Fig. 12. However, this integration has costs. If in terms of area, the penalty is negligible since only 19 sensors and 11 *CC* were inserted, while the power consumption may be more significantly impacted since *CC* cells toggle at each clock cycle. Similarly, the maximum operating frequency *Fmax* of the circuit might be slightly low due to the increase of the skew.

The consumption penalty is illustrated Table 10, comparing the back-end performance results of a typical flow (made in worst case PVT corner) versus our validation flow (made in nominal PVT

Table 9
Consumption and frequency performances in different PVT corners.

Consumption (mW)/ maximum frequency (MHz)	Typical flow	New flow
SS, 1.05 V, 125 °C	68/440	72/422
TT, 1.1 V, 25 °C	75/580	82/561
FF, 1.2 V, 125 °C	99/760	107/735

Table 8
Selecting a reduced set of critical endpoints.

	Number of worst endpoints obtained in nominal conditions remaining among the n ($n=1$, 10, 25, 50) worst endpoints obtained at other PVT conditions. Percentage of worst endpoints obtained in nominal conditions remaining among the n ($n=1$, 10, 25, 50) worst endpoints obtained at other PVT conditions. Worst endpoint captured (yes or no).					
	bc 1.2 V m40 °C	**bc 1.2 V 125 °C**	**nom 1.1 V 25 °C**	**wc 1.05 V 125 °C**	**wc 0.9 V 125 °C**	**wc 0.9 V m40 °C**
The worst endpoint ($n=1$)	1	1	1	1	0	0
	100%	100%	100%	100%	0%	0%
	yes	yes	yes	yes	no	no
The 10 worst endpoints ($n=10$)	8	8	10	8	7	7
	80%	80%	100%	80%	70%	70%
	yes	yes	yes	yes	yes	yes
The 25 worst endpoints ($n=25$)	22	21	25	23	22	22
	88%	84%	100%	92%	88%	88%
	yes	yes	yes	yes	yes	yes
The 50 worst endpoints ($n=50$)	46	45	50	45	45	45
	92%	90%	100%	90%	88%	88%
	yes	yes	yes	yes	yes	yes
Maximum frequency (MHz)	763	735	561	422	270	212
Minimum period (ns)	1.31	1.36	1.78	2.37	3.7	4.7
Greatest negative slack (ps) (10 worst endpoints)	56	68	60	100	150	190
Greatest negative slack (ps) (50 worst endpoints)	98	106	164	219	205	297

Table 10
Evolutions of the CT metrics.

		Prototype	After final CTS
Latency (ns)	**Typical case**	1.19	1.28
Skew (ps)		97	102
Levels		15	23
# buffers		846	258
Power at 440 MHz (mW)		22.8	18.8
Skew (ps)	**Best timing corner**	62	97
	worst timing corner	116	240

Table 11
Characteristics of the monitored test case block.

Technology	STMicroelectronics 45 nm low power
Test case	Arithmetic VLIW processor 32 bits
Area	0.338 mm^2
# Cells	54 k (Flip–flops: 13 k)
Frequency (1.05 V – 125 °C – SS)	422 MHz
Frequency (1.2 V – m40 °C – FF)	820 MHz
Power consumption (1.1 V – 25 °C – TT) at 440 MHz	82 mW
Static power (1.1 V – 25 °C – TT)	35 μW
# Monitored flip–flops	50
# CC cells	11
# S cells	19
# Added buffers in short paths	589

corner with relaxed paths). One can notice that the consumption penalties do not exceed 10%. One may also note that the effort to recover the loss of frequency due to the relaxed paths would not affect much the total power consumption. We will see next that the application of adaptive strategies will in an average compensate such an over-consumption. Table 11 summarizes all the final characteristics of the monitored block.

6.1.2. Behavioral verification

In order to check if the whole system works as planned, we extracted from the back-end level the maximum operating frequency *Fmax* of the circuit at different PVT conditions, as well as the maximum frequencies *Fmax_Sensor*(DW_{effi}) at which the monitoring system does not flag any timing warning or violations, considering, respectively, the two different detection windows: DW_{eff1} and DW_{eff2}. Table 12 gives the resulting simulated values of *Fmax_Sensor*(DW_{effi})/*Fmax* at different VT conditions for a typical process. The same kind of simulations has been done in worst and best process corners, demonstrating identical results (Table 13). For all obtained results, we observed that *Fmax_Sensor*(DW_{effi})/*Fmax* ratio remains below 100, meaning that the monitoring system operates correctly and warns that the circuit is close to a timing failure, in every PVT conditions.

In addition, the timing margin (difference [1/*Fmax*–1/*Fmax_Sensor* (DW_{eff1})]) ranges 80–340 ps. The use of a wider DW_{eff2} detection window implies the cost of a larger timing margin ranging 120–480 ps depending on the considered PVT conditions.

6.1.3. Performances

The results above demonstrate the efficiency of the monitoring system that may allow a circuit to work at a frequency roughly equal to 90% of its maximal effective speed under all PVT conditions. One may also conclude that it is interesting to integrate two detection windows (one introducing more timing margin than the other) to decide more efficiently at system level whenever lowering/increasing the frequency and scaling the voltage.

For better quantification of the differences between a typical flow and this work in the framework of an adaptive system, we simulated, in many different PVT conditions, our placed and routed block (denoted by standard design in the rest of the paper) that was designed using a worst case approach (typical flow) to operate at 430 MHz. We also simulated this block after an insertion of the monitoring system. Note that the resulting circuit, denoted by 'monitored design' below, has been synthesized considering nominal conditions, since the monitoring system aims at easing the integration of voltage and frequency scaling strategies. Figs. 14–16 give the evolutions of the power consumption of both designs w.r.t. the supply voltage but also the evolutions of the two periods 1/*Fmax* (standard design) and 1/*Fmax_Sensor*(DW_{eff1}) (monitored design) for a temperature of 125 °C.

For a nominal process (Fig. 14), the two considered designs consume the same amount of power under the same frequency

Table 12
Fmax_Sensor(DW_{effi})/*Fmax* at typical process.

Fmax_Sensor_*DW_{eff1}/Fmax*				
°C/V_{dd}	**0.9**	**1.0**	**1.1**	**1.2**
−40	95.0%	93.6%	93.9%	92.4%
25	92.7%	92.5%	92.7%	92.5%
80	91.8%	92.3%	90.5%	89.8%
125	92.1%	90.5%	90.3%	90.0%
Fmax_Sensor_*DW_{eff2}/Fmax*				
°C/V_{dd}	**0.9**	**1.0**	**1.1**	**1.2**
−40	92.1%	91.2%	90.6%	89.7%
25	89.8%	89.2%	89.4%	89.9%
80	89.3%	89.0%	87.4%	87.3%
125	88.8%	87.9%	87.2%	87.5%

Table 13
Fmax_Sensor(DW_{effi})/*Fmax* at worst (*SS*) and best (*FF*) processes, [0.8; 1.2] V, [−40, 125] °C.

	SS	**FF**
DW_{eff1}	[91–98]%	[87–93]%
DW_{eff2}	[85–90]%	[84–90]%

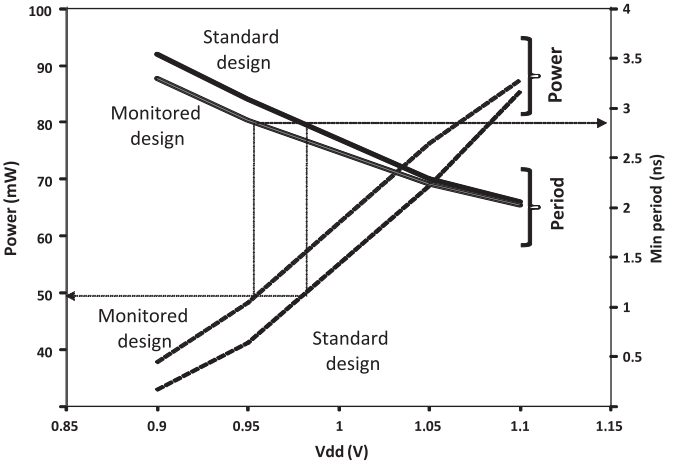

Fig. 14. Evolutions, for the nominal process, of the power consumption of both designs w.r.t. the supply voltage and evolutions of periods 1/*Fmax* (standard design) and 1/*Fmax_Sensor* (monitored design) at 125 °C.

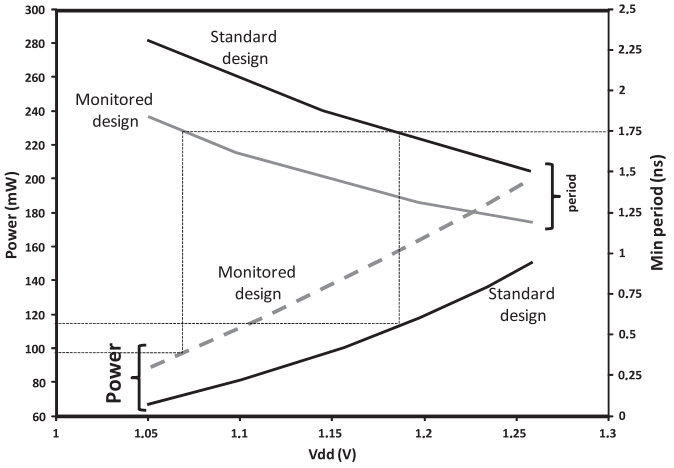

Fig. 15. Evolutions, for the best process, of the power consumption of both designs w.r.t. the supply voltage and evolutions of periods 1/*Fmax* (standard design) and 1/*Fmax_Sensor* (monitored design) at 125 °C.

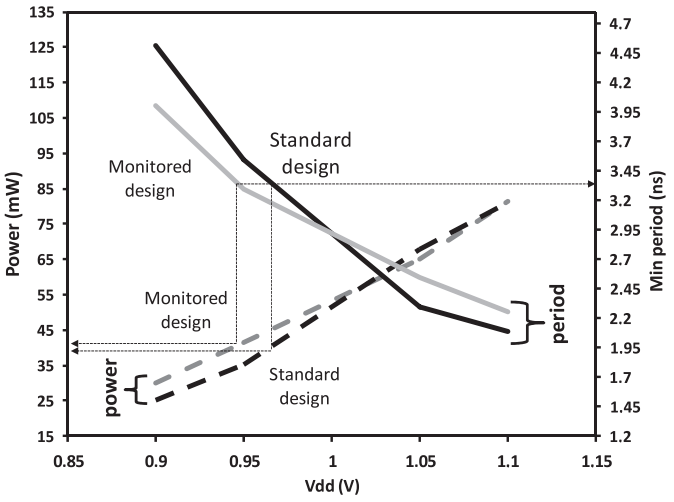

Fig. 16. Evolutions, for the worst process, of the power consumption of both designs w.r.t. the supply voltage and evolutions of periods 1/*Fmax* (standard design) and 1/*Fmax_Sensor* (monitored design) at 125 °C.

constraint; however note that the monitored design operates with a lower supply voltage due to the reduction of the timing margins.

For the best process (Fig. 15), it is possible, thanks to this monitoring system, to decrease significantly the supply voltage. The decrease may result, under timing constraint, in power savings as high as 20%. Better operating conditions (i.e. lower temperature value) can significantly further improve this gain.

These results are explained by the important timing margins observed if the fabrication is better than expected (best process): timing margins that can be traded for power, thanks to the monitoring system and voltage scaling. Note that, at production, such circuit might not be within the specifications due to its excessive power consumption. Thus, the proposed monitoring system may also help in improving the yield, as any post silicon tuning techniques do. Concerning now the slow process (Fig. 16), the timing margins are less important. Consequently the proposed system does not allow trading speed for power; and the additional hardware integrated can lead to an increase of up to 15% of the power consumption with our specific validation flow. However, during production, these circuits are significantly less consuming (two times less w.r.t. typical process in our case at the same operating frequency) and thus this penalty may not lead to a yield loss.

7. Conclusion

A new integrated sensor allowing monitoring the timing slack of critical paths has been introduced. Based on this sensor, a whole monitoring system allowing anticipating, at runtime and with a high timing resolution, timing slacks over a wide range of PVT conditions has been proposed. Its efficiency in monitoring timing slacks has been demonstrated on a 32 bits VLIW DSP, in 45 nm technology. Such a system may provide valuable and additional data to an adaptive system already integrating temperature, voltage and process sensors and is expected to provide power savings of up to 20% when used in combination with voltage scaling strategies.

References

[1] V. Migairou, R. Wilson, S. Engels, Z. Wu, N. Azemard, P. Maurine, "A simple statistical timing analysis flow and its application to timing margin evaluation", in: 16th International Workshop on Power and Timing Modeling, Optimization and Simulation PATMOS'07, Montpellier, France. September (2006) 138–147.
[2] S. Borkar, Designing reliable systems from unreliable components: the challenges of transistor variability and degradation, IEEE, Micro 25 (6) (2005) 10–16.
[3] D. Blaauw, K. Chopra, A. Srivastava, L. Scheffer, Statistical timing analysis: from basic principles to state of the art, IEEE Transactions on Computer-Aided design of Integrated Circuits and Systems 27 (4) (2008) 589–607.
[4] A. Davoodi, A. Srivastava, Variability-driven gate sizing for binning yield optimization, IEEE Transactions on Very Large Scale Integration Systems TVLSI'08 16 (6) (2008) 683–692.
[5] V. Narayanan, Z. Yan, E. Macii, S. Bhanja, in: Proceedings of the 18th ACM Great Lakes Symposium on VLSI 2008, Orlando, Florida, USA, May 4–6, ACM. (2008).
[6] B. Lasbouygues, R. Wilson, N. Azémard, P. Maurine, Temperature- and voltage-aware timing analysis, IEEE Transactions on Computer-Aided design of Integrated Circuits and Systems 26 (4) (2007) 801–815.
[7] C.R. Parthasarathy, A. Bravaix, C. Guérin, M. Denais, V. Huard, "Design-in reliability for 90–65 nm CMOS nodes submitted to hot-carriers and NBTI degradation", in: 16th International Workshop on Power and Timing Modeling, Optimization and Simulation PATMOS'07, Montpellier, France, September. 191–200.
[8] M. Nourani, A. Radhakrishnan, Testing on-die process variation in nanometer VLSI, IEEE Design & Test of Computers 23 (6) (2006) 38–451.
[9] S.B. Samaan, "Parameter variation probing technique", US Patent 6535013. (2003).
[10] M. Persun, "Method and apparatus for measuring relative, within-die leakage current and/or providing a temperature variation profile using a leakage inverter and ring oscillators", US Patent 7193427. (2007).
[11] H.-J. Lee, "Semiconductor device with speed binning test circuit and test method thereof", USPatent 7260754. (2007).
[12] Z. Abuhamdeh, B. Hannagan, Jeff Remmers, L.Crouch Alfred, A production IR-drop screen on a chip, IEEE Design & Test of Computers 24 (3) (2007) 216–224.
[13] A. Drake, R. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, V. Pokala, A distributed critical path timing monitor for a 65 nm high performance microprocessor, IEEE Journal of Solid-State Circuits (2007) 398–399.
[14] S. Das, P. Sanjay, D. Roberts, S.W. Lee, D. Blaauw, T. Austin, T. Mudge, K. Flautner, A self-tuning DVS processor using delay-error detection and correction, IEEE Journal of Solid-State Circuits (JSSC) 41 (4) (2006) 792–804.
[15] S. Das, C. Tokunaga, S. Pant, W.H. Ma, S. Kalaiselvan, K. Lai, D.M. Bull, D.T. Blaauw, Razor II: in situ error detection and correction for PVT and SER tolerance, IEEE Journal of Solid-State Circuits (JSSC) 44 (1) (2009) 32–48.
[16] K.A. Bowman, J.W. Tschanz, J.C. Nam Sung Kim Lee, C.B. Wilkerson, S.-L.L. Lu, T. Karnik, V.K. De,"Energy-efficient and metastability-immune timing-error detection and instruction-replay-based recovery circuits for dynamic-variation tolerance", International Solid-State Circuits Conference Digest Technical Papers (ISSCC'08), Francisco CA, USA, 402–623. February (2008).
[17] Satish Yada, Bharadwaj Amrutur, Rubin A. Parekhji, "Modified stability checking for on-line error detection", in: 20th International Conference on VLSI Design, VLSID'07, Bangalore, India. February 2007 787–792.
[18] C. Bernard F. Clermidy, "A low-power VLIW processor for 3GPP-LTE complex numbers processing", Design, Automation and Test in Europe, DATE'11, 14–18 March, Grenoble, France. 2011.
[19] T. Nakura, K. Nose et M. Mizuno, "Fine-grain redundant logic using defect-prediction flip-flops", International Solid-State Circuits Conference Digest Technical Papers (ISSCC'07), San Francisco CA, USA. February 2007, 402–403.
[20] M. Agarwal, B.C. Paul, M. Zhang, S. Mitra., "Circuit failure prediction and its application to transistor aging", in: Proceedings of the 25th VLSI Test Symposium, VTS'07, Berkeley CA, USA. May (2007) 277–286.
[21] T. Kehl, "Hardware self-tuning and circuit performance monitoring", in: 1993 International Conference on Computer Design (ICCD-93). October (1993) 188–192.