Monitor Circuits for Cross-Layer Resiliency

Mahfuzul Islam and Hidetoshi Onodera

1 Introduction

The end of supply voltage scaling has pushed circuit designers to find for new solutions to reduce power consumption. One key reason for the stoppage of supply scaling is variability including aging. The International Technology Roadmap for Semiconductors (ITRS) highlights performance variability and reliability management in the next decade as a red brick (i.e., a problem with no known solutions) for the design of computing hardware [1]. Instead of operating under predefined supply voltage and clock frequency, the circuit must adapt itself according to its process conditions, as well as to the dynamic changes of temperature, aging, and workload to harness the full potential of technology scaling. With the resilient operations, a chip's lifetime can be extended, and energy consumption can be reduced.

Due to significant variations in temperature, workload, and aging, dynamic tuning of not only the supply voltage and clock frequency but also the threshold voltages has become a necessity for energy-efficient operation. However, without knowing the device and environmental parameters, tuning of these parameters is not possible. On-chip monitor circuits which provide the information about device and environment come to play an important role. On-chip monitors realize an interface between hardware and software, which then can be utilized for software-controlled optimization. The future LSI (Large Scale Integration) chip will require lots of monitors to track transistor performances, temperature changes, supply voltage droops, and leakage current variations. This chapter describes some design techniques of monitor circuits based on delay cells and then presents a reconfigurable monitor

M. Islam (✉) · H. Onodera
Kyoto University, Kyoto, Japan
e-mail: islam.akmmahfuzul.3w@kyoto-u.ac.jp; onodera.hidetoshi.4x@kyoto-u.ac.jp

architecture to realize different delay characteristics with a small area footprint. An extraction methodology of physical parameters from a set of monitor circuits is presented for model-hardware correlation.

2 Cross-Layer Resiliency

This section describes the benefit of realizing cross-layer resiliency by dynamic tuning of threshold voltage, supply voltage, and clock frequency. Cross-layer resiliency enables energy-efficient operation by eliminating excessive margins. We highlight the importance of run-time sensing of circuit delay, leakage current, switching power, temperature, and threshold voltage to realize minimum energy operation under process, voltage, temperature, activity and temperature variations. Multiple on-chip monitor circuits are required to sense these parameters. Although monitor circuits are not a part of the actual circuit, they are essential components for run-time tuning.

2.1 *Parameter Fluctuation and Aging*

Variations in physical parameters such as transistor threshold voltage, and temperature have spatial distributions over a chip with both of the random and systematic components. Besides the physical parameter variations, environmental variations also affect circuit performance significantly. Temperature variations of more than 50 °C between different parts within a chip are reported [2]. Increase in temperature degrades circuit performance and increases leakage power. According to ITRS, supply voltage fluctuation is considered to be $\pm 10\%$ of the nominal voltage. Sudden drop of supply voltage may cause critical timing failure causing system malfunctioning. Because of process variation, some chips can be slow and some chips can be fast. Fast chips tend to be leaky causing larger energy consumption. Designers thus face a challenge to meet both of the delay and power constraints, since the circuit needs to operate correctly under all of the variation scenarios.

Device characteristics also degrade over time. Aging causes reliability issues where high temperature accelerates device aging. Device phenomena such as Negative Bias Temperature Instability (NBTI) is reported to cause 10% of delay degradation in digital circuits for a 70-nm process over 10 years [3]. Designing the circuit for the worst possible scenario is energy inefficient as it increases area, power, and cost. A chip may face extreme worst-case scenarios once in several years. The conventional worst-case design methodology, where the operating conditions of a circuit are set such as to meet the worst-case performance, is way too energy inefficient and new design paradigm incorporating on-chip monitor circuits have become indispensable. In the new design paradigm, parameters such as the supply voltage and threshold voltage are tuned in the run-time such that the target delay

and power profile are achieved. As a result, instead of worrying for the worst-case performance, the circuit can now be designed to achieve optimal performances.

2.2 Cross-Layer Resiliency for Energy-Efficient Operation

Figure 1 shows a typical design hierarchy of a system-on-a-chip. First, transistor models for a target process technology node are given to circuit designers. These transistor models contain statistical models to simulate the effects of variations on circuit performance. To guarantee error-free circuit operation, a circuit is tested for extreme cases by using the assumed models. As a result, the circuits tend to be over-designed which result in excessive energy consumption. From a system perspective, the circuits need to operate at different supply voltages and clock frequencies while ensuring correct operations. The selection of adequate clock frequency and supply voltage is performed pessimistically. Design-time optimization is an open-loop operation; thus the operating conditions are set for the worst-cases. The solution obviously is to create a feedback loop into the system which can only be realized by tuning circuit parameters in the run-time. Run-time tuning relaxes the design constraints on the circuit and as a result the circuit become better optimized compared with the one where no run-time tuning is performed.

Figures 2 and 3 show two profiles of energy consumption for an LSI. Figure 2 shows simulated energy and frequency contour plots on the threshold voltage (V_{th}) and the supply voltage (V_{dd}) plane for a model circuit operating at an activity rate of 1%. The model circuit used here is a delay line of 40 inverter cells. A commercial

Fig. 1 Cross-layer optimization with the use of monitor circuits

Fig. 2 Energy and frequency contour plot on the V_{th} and V_{dd} plane. Activity rate of 0.01 is assumed

Fig. 3 Total energy per clock cycle against the ratio between static and dynamic energy for a clock frequency of 100 MHz. Having a balanced static and dynamic energy is the key to minimum energy operation

65 nm process is assumed here. Cross points in the plot show the sets of V_{th} and V_{dd} values that give the minimum energy operation for each operating frequency. We observe that the required V_{th} and V_{dd} values, that realize the minimum energy operation, differ significantly with the changes in the clock frequency. Dynamic adaptation of V_{th} and V_{dd} values ensures minimum energy operation for any operating frequency. Figure 3 shows the total energy of the circuit operating at 100 MHz under different combinations of V_{th} and V_{dd} against the ratio of static energy (E_{static}) to dynamic energy ($E_{dynamic}$). We observe that a ratio of 10 to 50% realizes near minimum energy operation. Under the variations of circuit activity, operating frequency and temperature, the energy ratio varies largely. To ensure minimum energy operation, V_{dd} and V_{th} values need to be tuned such that a ratio between 10 and 50% is realized. From the figures, the need for run-time tuning of V_{dd} and V_{th} values are apparent but the problem is how to realize such a mechanism.

Two key mechanisms are required to realize a feedback system. One is the sensing mechanism of the output. The other is to feedback the output to the input of the system. Sensing mechanism is an essential component here. In the case of an LSI, the output parameters are the V_{th} values, circuit delays, temperature, leakage current, and switching current. Sensing these parameters requires multiple on-chip monitor circuits. The monitors provide real-time information of the hardware which can then be used to set the parameters of V_{dd} , V_{th} and clock frequency optimally for reliable operation.

2.3 *Role of Monitor Circuits*

The past trend of using smaller transistors to achieve higher operating frequency has come to an end [4]. Instead of the clock frequency, system throughput and energy per throughput are the modern specifications for a device. The new era of LSI scaling is a system-on-a-chip (SoC) approach that combines a diverse set of components including adaptive circuits, integrated on-chip monitors, sophisticated power-management techniques, and increased parallelism to build products that are many-core, multi-core, and multi-function [5]. The ability to adapt to the changes in environment and performance will give us the full benefit of technology scaling. Tuning mechanisms and on-chip monitors are needed to realize circuits that have the ability to adapt. The future SoC must have capabilities of post-silicon self-healing, self-configuration, and error correction. Effective use of on-chip monitor circuits will play a major role in continuing the advancement of LSI. Use of on-chip monitors provides us the following advantages:

1. Reduce design margin in each layer of design hierarchy by eliminating pessimism.
2. Tune system parameters based on the actual hardware profile.
3. Provide information for silicon debugging and timing analysis.

To harness the above advantages, the following characteristics of on-chip monitor circuits are preferred:

Digital	Digital in nature realizes robust operation under different supply voltages.
Design automation	Monitor circuits for threshold voltage, temperature, supply voltage, interconnect, activity, and leakage current are required. Thus, design automation is a key factor here for low-cost implementation of the monitors. Cell-based design with delay cells are preferred.
Area efficiency	Area efficiency is an important parameter for fine-grain and distributed implementation of monitor circuits on the chip.

As the target parameters such as the temperature and leakage current are analog values, mechanisms to convert the analog values to digital values are required to interface with the other components of the system. Two design methodologies can be adopted for designing monitor circuit. One methodology performs operations in the analog domain to sense and amplify the effect of the parameter variation and then convert the analog value to a digital value. The other methodology converts the analog value to a digital value as early as possible and then make operations in the digital domain. Incorporating the analog value in the delay of a logic gate realizes the later. Furthermore, the well established cell-based design methodology for automation can be adopted readily for the delay-based implementation of monitor circuits. We therefore explore several delay-based implementations of monitor circuits in this chapter.

3 Delay-Based On-Chip Monitor Design

Delay-based monitor circuits use the mechanisms of converting the target analog value to the delay of a logic gate. The topology of the logic gate thus need to be designed such that the target parameter variation is amplified in the delay. To understand the delay-based monitoring, we first give an overview of the general delay characteristics of logic gates. Then we explore several techniques to tune the delay characteristics such that the monitoring of a target parameter can be realized. Finally, we demonstrate a cell-based design of a reconfigurable monitor circuit that can sense the parameters of nMOSFET and pMOSFET threshold voltages.

3.1 Delay Characteristics

Delay-based monitoring is based on the fact that the delay of a logic gate contains information of the transistor drain current I_d . Figure 4 shows four delay paths consisting of different logic gates and interconnects. A delay path of Fig. 4a consists of inverter gates. Delay paths of Fig. 4b and c consist of NAND2 and NOR2 gates. A delay path of Fig. 4d consists of inverter gates with long interconnecting wires. Depending on the topology of the logic gate and the interconnect length, delays of different gates and interconnect show different behavior to process, voltage, and temperature variation. Figure 5 shows the topology of four different logic gates. Figure 5a shows a conventional inverter topology. Figure 5b shows a NAND2 topology where two nMOSFETs are placed in stack. Figure 5c shows a NOR2 topology where two pMOSFETs are placed in stack. Figure 5d shows an inverter topology where two pMOSFETs and two nMOSFETs are placed in stack to mimic the delay behavior of the both of the NAND2 and NOR2 gates.

Under the presence of large within-die random variation, each delay path might behave differently. At a higher supply voltage, a particular path may show the

Fig. 4 Delay paths consisting of (a) inverter gates, (b) NAND2 gates, (c) NOR2 gates, and (d) inverter gates with long wires

Fig. 5 Topology of different delay cells. (a) Inverter gate. (b) NAND2 gate. (c) NOR2 gate. (d) Universal delay cell

worst-case delay, whereas at a lower supply voltage, a different path may show the worst-case delay. Figure 6 shows the delay change against the change of supply voltage. Topology with a stacked transistor shows higher sensitivity to V_{dd} change than that without a stacked transistor. Topology with a reduced V_{gs} value shows much higher sensitivity to V_{dd} change. The important point is that the delays of different topologies show different sensitivities to process, supply voltage, and temperature changes. Under the presence of within-die variation, the gates of the same logic type also show different delay behavior. Thus, accurate delay estimation of a circuit is challenging. Instead, we can monitor the delay of a representative circuit that gives us a reasonable prediction of the actual delay of the circuit.

Fig. 6 Delay versus supply voltage for different inverter topologies

3.2 Delay Model

A delay model is useful to intuitively understand the different delay characteristics for different topology. The rise and fall delays of an inverter gate can be approximated by the following equations:

$$d_{\text{rise}} = \frac{C_{\text{load}} V_{\text{logic}}}{I_{\text{dp}}}, \quad (1)$$

$$d_{\text{fall}} = \frac{C_{\text{load}} (V_{\text{dd}} - V_{\text{logic}})}{I_{\text{dn}}}. \quad (2)$$

Here, I_{dp} and I_{dn} are the drain currents of pMOSFET and nMOSFET during the ON state, respectively. C_{load} is the load capacitance that consists of the gate capacitance of MOSFETs of the next gate, drain capacitance of pMOSFET and nMOSFET, and interconnect parasitic capacitance. V_{logic} is the logical threshold voltage at which the next gate switches its output value. To model the transistor drain current, EKV model based equation of Eq. 3 is useful to express the drain current that is continuous from weak-inversion to strong-inversion operation: [6, 7].

$$I_d = k \cdot \frac{W}{L} \cdot \ln^\alpha \left[1 + \exp \left(\frac{V_{\text{gs}} - (V_{\text{th}} - \gamma V_{\text{bs}} - \lambda V_{\text{ds}})}{\alpha n V_T} \right) \right]. \quad (3)$$

Here, k is a technology-related parameter. γ is the body bias coefficient and λ is the short-channel coefficient. Short-channel effect reduces the threshold voltage when large V_{ds} is applied to the transistor. Thus, large V_{ds} value increases ON current which is beneficial to switching delay, but causes exponential increase in the leakage current.

For the pull-down operation of an inverter gate of Fig. 5a, V_{bs} is zero and V_{ds} changes from V_{dd} to V_{logic} . However, in the case of a NAND2 gate, the values of V_{bs} and V_{ds} differ. The source of the nMOSFET that is connected to the output is

not tied to ground. As a result, V_{bs} becomes negative that causes the V_{th} to increase. Consequently, the V_{ds} value remains within a small value. Smaller V_{ds} value causes less short-channel effect resulting in a higher V_{th} value than a larger V_{ds} value. As a result of negative V_{bs} value and smaller V_{ds} value, the drain current decreases which causes the delay to increase.

3.3 Delay-Based Monitor Circuits

Design of on-chip monitors requires careful choosing of the right topology. Here, we discuss several delay-based design techniques that realize monitoring of different parameters.

3.3.1 Critical Path Monitor

The first and the most important parameter to monitor is the maximum delay of a circuit to ensure that the circuit operates at a certain clock frequency without any timing error. The maximum delay of a circuit is the maximum of delays of all the paths. As a circuit consists of thousands of delay paths, we can choose the following two methods to monitor the maximum delay.

1. Monitor the delays of actual paths, and
2. Monitor the delay of a representative delay path.

The first method, which is in-situ monitoring, requires additional circuitry in the actual delay paths. In the case of in-situ monitors, the Flip-Flops (FF) in a circuit are replaced with special FFs with error detection sequential (EDS) functions. The EDS can either detect whether a timing error has occurred [8, 9] or warn us before the occurrence of actual errors [10–12]. Supply voltage and clock frequency are adapted accordingly based on the EDS signals. The drawback of EDS-based in-situ monitors is that the additional circuits add extra delays, and increase area and power. To reduce the delay and area overhead, we can replace only those FFs where the delays are critical. During the design phase, we can make a list of the potential critical delay paths. However, as shown in Fig. 6, paths show different sensitivity to process, supply and temperature changes. Thus, the number of candidates tend to increase drastically under process, voltage, and temperature variations. Another fundamental drawback to be overcome is that a critical path is not always sensitized. Thus, it is necessary to properly estimate the actual timing slack of the critical path.

The second method requires an additional delay path that is placed near the actual circuit that can track the delay of the actual circuit. This delay path is often called a critical path monitor (CPM). The requirement of such a CPM is that it tracks the maximum delay of the target circuit for all conditions of process, voltage, and temperature variations. CPM is thus a delay path that is synthesized such that it tracks the worst delay of the circuit. However, there is no universal solution on how

Fig. 7 Synthesis of critical delay path from a combination of series and parallel delay paths. (a) Parallel paths. (b) Series paths

to design a CPM that meets the above criterion. Two approaches have been proposed on how to synthesize a CPM. One approach is to synthesize a critical path monitor from a list of potential critical paths during the design phase [13–15]. The other approach is to design a reconfigurable delay path consisting of different logic gates and wire lengths, and then configuring the delay path during the test time, such that the delay correlates with the maximum achievable frequency [16–19]. Figure 7 shows a general concept of the synthesis framework of a critical delay path [20]. Several paths such as the paths shown in Fig. 4 are put in parallel. Then the several paths are placed in series. During the calibration process, combinations of parallel and series paths are explored to find a combination that gives the worst delay for all the operating conditions.

Instead of using a reconfigurable delay line, a general purpose delay line consisting of inverter cells with stacked transistors are also proposed so that the path mimics the worst-case delay [21]. Calibration is nonetheless required which can be performed during the design phase and during the test. To encounter the effect of systematic within-die variations, multiple CPMs can be used that are distributed at various places on the chip [15, 21].

3.3.2 Threshold Voltage Monitor

For adaptation of V_{th} values to their optimum values, V_{th} monitors are required. Although there is no universal definition of V_{th} , an arbitrary definition can be used as a reference. For example, the V_{gs} value that gives a fixed I_d value is often used to define the V_{th} value. Conversely, we can track the V_{th} value by observing the change of I_d value if the V_{gs} can be set as a function of V_{th} . Then the delay change resulting from the I_d change can be measured and converted to digital with the use of a reference clock signal. Figures 8 and 9 show two delay cells consisting of inverter gates where either the nMOSFET V_{gs} or the pMOSFET V_{gs} voltage becomes a function of the corresponding V_{th} values (V_{thp} for pMOSFET and V_{thn} for

Fig. 8 V_{thp} -dominant delay cell for V_{thp} monitoring

Fig. 9 V_{thn} -dominant delay cell for V_{thn} monitoring

nMOSFET). The V_{th} -sensitive gate-source voltage is realized using pass-transistors as shown in Figs. 8 and 9 [22, 23]. To illustrate the V_{th} monitoring capability of the monitor cells, sensitivity vectors of different inverter topologies are shown in Fig. 10 at nominal supply voltage for a 65 nm bulk process. Here, the sensitivity vector consists of the sensitivity coefficients of the delay to V_{thn} and V_{thp} changes. We observe that the sensitivity coefficients of the pass-transistor inserted cells are multiple times larger than those of conventional inverter, NAND2, and NOR2 cells. An all-digital process variability monitor based on a shared structure of a buffer ring and a ring oscillator is proposed in [24]. The technique utilizes the differences of rise and fall delays of inverter gates because of process variations.

As will be shown next, driving the load with the transistor leakage current also gives us a delay that is exponentially related to V_{th} value change. However, driving the load using leakage current requires careful design because leakage currents through the pull-up and the pull-down paths get involved also. Gate-leakage current is also a factor to degrade the accuracy of such monitors. The topologies of Figs. 8 and 9 give us compact designs that are minimal and fulfill the purpose.

Fig. 10 Sensitivity vectors of different inverter topologies

Fig. 11 I_{offp} -dominated delay cell

3.3.3 Aging Monitor

A critical path monitor also acts as an aging monitor. However, the differences in activity rate may cause deviations in the aging between an actual critical path and a monitor path. Therefore, delay paths of different activity rates can be implemented to track aging. Multiple delay lines consisting of inverter gates with different activity rates would give us precise aging information. Decoupling the NBTI and PBTI effects can be useful for debugging and modeling purposes. In that case, different architectures are proposed for independent NBTI and PBTI monitoring [25].

3.3.4 Sub-threshold Leakage Monitor

Sub-threshold leakage monitor helps us to estimate the leakage current of a circuit. The information can then be used to tune V_{th} , V_{dd} or frequency optimally. Figure 11 shows a delay cell whose rise delay is several orders of magnitude larger than the fall delay. The rise delay is driven by the pMOSFET OFF current, while the fall

Fig. 12 I_{offn} -dominated delay cell

Fig. 13 Temperature monitoring utilizing inverter delay driven by nMOSFET OFF current. (a) Logarithm of oscillation period driven by nMOSFET OFF current against temperature. (b) Monitoring error against temperature after an one-point calibration

delay is driven by the nMOSFET ON current. As a result, the delay of a path consisting of this cell is proportional to the pMOSFET OFF current. Similarly, delay of a path consisting of delay cells of Fig. 12 is proportional to the nMOSFET OFF current. Figure 13a shows the change of measured oscillation period for a delay path consisting of 125 inverter stages against the temperature change. The inverter topology of Fig. 12 is used here. The target process is a 65 nm bulk process. The oscillation period here corresponds to the average OFF current of 125 nMOSFETs. The logarithm of the delay changes linearly with the temperature showing that the monitor tracks the leakage current change correctly.

3.3.5 Temperature Monitor

As leakage current is sensitive to temperature variation, a leakage current monitor can be used for on-chip temperature monitoring. The logarithm of the oscillation period, D , can be expressed by the following equation:

$$\ln(D) = a_T + b_T \cdot T, \quad (4)$$

where T is the absolute temperature, a_T and b_T are temperature coefficients. Figure 13b shows the monitoring error after a one-point calibration for a 65 nm bulk process. Calibration is performed at 15 °C. An error range of -1.3 °C to 1.4 °C is observed. The above error range is small enough for real-time thermal and reliability management.

3.3.6 Supply Voltage Monitor

Supply voltage fluctuation has always been a concern which is getting more severe with the reduction of supply voltage. Supply voltage fluctuation has a static component which results from the power delivery network (PDN) and a dynamic component which is the result of transition from idle to active state of a circuit. As critical path monitors are also sensitive to supply voltage fluctuations and have a high bandwidth, they can also detect dynamic supply voltage fluctuations [17]. In the case of CPMs, the output is the timing information obtained by comparing the path delay and clock period. Thus, the error information does not give whether the error is from temperature or supply voltage for example. However, when combined with other monitors such as temperature and threshold voltage, identification of the causes of timing error becomes possible. The identification of the sources of timing error allows correct optimization and lifetime enhancement. On-chip supply voltage droop monitoring mechanisms have been proposed to evaluate the power delivery network (PDN) [26].

3.3.7 Activity Monitor

Run-time estimation of the static and the dynamic energy can be used to achieve the minimum energy operation as suggested by Fig. 3. As the dynamic energy is proportional to circuit activity rate, we can estimate the dynamic energy by calculating the activity rate of a circuit. A digital dynamic power meter (DDPM) has been used that computes a rolling average of signal activity over a fixed number of clock cycles [27]. The accuracy of the power estimation here depends on careful selection of signals, such that they correspond to the activity of structures that have high power consumption. Instead of monitoring key logic signals, a clock activity adder (CAA) for switching power estimation is also proposed [28]. The approach of the CAA takes advantage of the fact that switching power is highly correlated to register clock activity. Similarly, hardware-event monitors such as memory-access counters and instruction-execution counters can be used for dynamic energy estimation [29]. These monitors depend on counting signal transitions rather than the delay itself.

Fig. 14 A reconfigurable inverter cell topology for V_{thp} and V_{thn} monitoring. “C” is a control signal. (a) Reconfigurable topology. (b) V_{thp} -sensitive configuration, and (c) V_{thn} -sensitive configuration

3.4 Reconfigurable Delay Path for Multiple Parameter Monitoring

Delay-based sensing enables us to design a reconfigurable architecture to monitor multiple parameters by configuring the delay path accordingly [30, 31]. For example, we can use the topology of Fig. 14a to monitor both of the V_{thp} and V_{thn} variations. Figure 14b and c shows the two configurations to make the delay V_{thp} - and V_{thn} -sensitive, respectively.

3.5 Cell-Based Design

The use of delay cells provides the advantage of the use of cell-based design flow that enables us to place and distribute the monitors into different parts of the chip. For example, temperature monitors need to be placed at hot-spots where power density is high. Power density maps are generated during the design phase. A cell-based design example in a 65 nm bulk triple-well process for a reconfigurable V_{th} monitoring circuit is shown in Fig. 15. The cells with green highlights in Fig. 15b are the monitor cells of Fig. 15a. The placements of the cells are performed carefully utilizing the “do not touch” and “relative adjacent placement” features of the place and route tool.

3.6 On-Chip Measurement and System Interface

The monitoring circuit needs to be interfaced with system for adaptation and self-tuning. The following three mechanisms can be adopted for on-chip measurement of monitor circuits.

Fig. 15 Delay characteristics for different topology and supply voltage. (a) Cell layout of a reconfigurable V_{th} monitor delay cell. (b) Chip micrograph and layout of a reconfigurable monitor circuit including the controller

Fig. 16 Three different measurement methods for system interfacing. (a) EDS-based method. (b) Time-to-digital conversion based method, and (c) Frequency ratio based method

1. Edge detection [16, 17, 32].
2. Frequency counting [33].

Edge detection based system can have either a single bit output [16] or multiple bits output [17, 32]. Figure 16 shows three different methods for digitizing the monitored delay. Figure 16a checks whether the delay is smaller or larger than the system clock period [17, 32]. If the delay is smaller, adaptation such as slowing

down the system by reducing the supply voltage can be performed. If the delay is larger, the system will speed up by increasing the supply voltage for example. To ensure that the transitions occur without any timing error, margins are added in the delay path. These margins include within-die random delay effects as well as the response time of the adaptation. A resolution window can also be added to ensure that the adaptation occurs without inducing any timing error.

Figure 16b uses multiple edge detectors to convert the time between the path delay and the clock period to digital codes [16]. The digital codes are then sent to the system controller where a look-up table (LUT) based adaptation can be implemented. Figure 16c shows a measurement method that uses frequency counting [33]. This measurement method is particularly useful for monitoring device parameters of V_{th} , temperature, and so on. Using the system clock for the conversion will require calibration of the monitoring circuit for every supply voltage which will increase the test cost. Instead, we can utilize a locally generated clock using a ring oscillator. The output in this case is the ratio of the monitor frequency and the reference frequency. The measured values of the frequency ratio are then compared with predefined values to monitor how much the monitoring parameter varied from the targeted values. For applications where the clock frequency is fixed, process and temperature sensitive monitors can also be implemented with edge detection mechanisms. An up/down counter based detection circuit to detect the V_{th} deviation from predefined values has been employed for dynamic adaptation of V_{th} values [34].

4 Parameter Extraction for Model-Hardware Correlation

The circuit techniques described in Sect. 3 realize delay characteristics that are sensitive to particular parameter variations. However, they do not give us the value of the parameter variation itself. In this section, we describe a parameter extraction technique that takes the delay values of multiple delay paths and then estimates the variations in each of the parameters. The parameters can be transistor threshold voltage, temperature, gate-length or any device related parameter. We can then utilize the extracted parameters for test strategies and process optimization.

4.1 Parameter Extraction Methodology

In the case of an inverter gate, the gate-source voltage of each transistor goes through different values during a “High” to “Low” and a “Low” to “High” switching events. Thus, it is not straight forward to relate physical device parameters to the delay information. Parameter estimation gets harder when the supply voltage is lowered, such that the delay becomes non-linear to the parameter changes. So, the question is how to correlate the model to each chip to get a good accuracy.

Fig. 17 Estimation of physical parameters from multiple delay paths. Sensitivity coefficient links the physical parameters to delay values

Section 3 demonstrates different types of delay paths to monitor various physical and environmental parameters. The designs are carefully performed to make the delay particularly sensitive to the parameter of interest. The techniques allow us to comparatively track the change of the parameters in the run-time. However, because of the mismatch in the model and hardware, the absolute parameter monitoring contains errors. Calibrations need to be performed to reduce the errors to acceptable ranges. For debugging purposes, we may want to correlate our transistor models to actual transistor characteristics in the chip. To perform model-hardware correlation, key model parameters such as the V_{th} and β may suffice as they are the dominant sources of fluctuations, although other parameters may also be used.

Figure 17 illustrates the concept of parameter extraction from multiple delay values. The left side of the figure plots the delay of a path against the delay of a different path. The round point shows a point which is obtained by circuit simulation. The cross point emulates a measured value from a chip. The difference between the two points here contains process information. Using sensitivity coefficients, we can estimate the amount of deviation in the process parameters and transform the delay space to process space which is shown in the right side of the figure. The key point here is not to use transistor I–V characteristics, rather use the delay characteristics to extract these parameters. For robust extraction of the parameters, we need to design the delay paths, such that the sensitivity matrix has a low condition number [22]. We can then build a system of linear equations using the sensitivity coefficients.

4.2 Measurement Results

To demonstrate the monitoring capability of the V_{thp} -sensitive and V_{thn} -sensitive delay cells of Figs. 8 and 9, measurements of ring oscillators are performed for a 65 nm bulk process. Figure 18 plots the values of V_{thp} and V_{thn} estimated under different body bias conditions for a particular chip. In the figure, the x-axis refers to V_{thp} estimations and the y-axis refers to V_{thn} estimations. Rectangular points are

Fig. 18 Estimation results of V_{thn} and V_{thp} of a chip for different body bias values. Either the pMOSFETs or the nMOSFETs are biased simultaneously

Fig. 19 V_{thp} -sensitive RO (ring oscillator) frequencies against V_{thn} -sensitive RO frequencies

estimated values of V_{thp} and V_{thn} when only pMOSFET is biased. Triangular points refer to estimated values of V_{thp} and V_{thn} when only nMOSFET is biased. When only pMOSFET is biased, the estimated point moves in the horizontal direction referring that only V_{thp} is being changed in the estimation. When only nMOSFET is biased, the estimated point moves in the vertical direction referring that only V_{thn} is being changed in the estimation. Thus, it is demonstrated that any change in the threshold voltage can be detected correctly by the proposed monitor circuits.

Figure 19 shows the measured frequencies of V_{thp} -sensitive and V_{thn} -sensitive ring oscillators from several chips (open circles). The chips have been fabricated targeting either of the five process corners of “TT,” “SS,” “FF,” “FS,” and “SF.” The values are normalized by the values simulated with the transistor models targeted for the “TT” process corner. Frequency values simulated using the other corner models

Fig. 20 V_{th} estimation results for different corner chips

are also plotted in the figure (closed squares). In the figure, process shifts from the “TT” model prediction are observed. Clear deviations are observed for “TT,” “SS,” “SF,” and “FS” corners. The silicon values are higher than the model predictions. With comparison with the models, we can have quick understanding of process shift for each chip. This information allow us to take decisions for silicon debug and test pattern generation. We can now extract the device parameters of V_{thp} , V_{thn} , and β using sensitivity analysis, model-hardware correlation can be obtained that allows us accurately predict the delay performance. Figure 20 plots the estimated V_{thp} and V_{thn} values. V_{th} values provided in the corner models are also plotted in the figure. Furthermore, V_{th} values provided by the Process Control Modules (PCM) that are generally placed in the scribe-lines are also plotted. The estimated values correlate with the PCM data and also show die-to-die variations.

5 Conclusion

In this chapter, we have shown the importance of cross-layer resiliency for energy-efficient and robust operation of circuits. Cross-layer resiliency is performed by tuning the threshold voltage and supply voltage in run-time based on information of process, leakage current, circuit activity, and temperature. Run-time monitoring of these parameters are essential in achieving cross-layer resiliency. To incorporate the monitor circuits into a cell-based design flow, we have discussed delay-based monitoring techniques. Cell-based design of monitor circuits enables to place the monitors inside the circuit. Placing the monitors inside the target circuit realizes better correlations between the monitor behavior and the actual circuit behavior.

We have discussed a general design methodology to synthesize a critical path monitor. There are several methods to monitoring the critical delay having a trade-off relationship between accuracy and implementation cost. Implementation cost

here can be area overhead, test cost, and/or both. The selection of a suitable critical path monitor thus has to be made based on the critical nature of the application.

Besides the critical path monitoring, run-time monitoring of physical parameters of threshold voltage, temperature, and leakage current are essential for energy-efficient operation under parameter fluctuation and aging. Utilizing the relationship between the delay of a logic gate and the physical parameters, several circuit topologies are discussed that amplify the effect a certain parameter. Threshold-dominant inverter topologies and leakage current driven inverters are suitable for temperature, leakage current, and threshold voltage monitoring.

References

1. ITRS, International Technology Roadmap for Semiconductors. Technical Report [Online]. Available: <http://www.itrs.net>
2. Borkar, S., Karnik, T., Narendra, S., Tschanz, J., Keshavarzi, A., De, V.: Parameter variations and impact on circuits and microarchitecture. In: Design Automation Conference, pp. 338–342 (2003)
3. Paul, B.C., Kang, K., Kufluoglu, H., Alam, M.A., Roy, K.: Impact of NBTI on the temporal performance degradation of digital circuits. *IEEE Electron Device Lett.* **26**(8), 560–562 (2005)
4. Horowitz, M.: 1.1 Computing's energy problem (and what we can do about it). In: IEEE International Solid-State Circuits Conference, pp. 10–14 (2014)
5. Bohr, M.: The new era of scaling in an SoC world. In: IEEE International Solid State Circuits Conference, pp. 23–28 (2009)
6. Enz, C.C., Krummenacher, F., Vittoz, E.A.: An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications. *Analog Integr. Circ. Sig. Process* **8**(1), 83–114 (1995)
7. Marr, B., Degnan, B., Hasler, P., Anderson, D.: Scaling energy per operation via an asynchronous pipeline. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **21**(1), 147–151 (2013)
8. Ernst, D., Kim, N.S., Das, S., Pant, S., Rao, R., Pham, T., Ziesler, C., Blaauw, D., Austin, T., Flautner, K., Mudge, T., Ave, B., Arbor, A.: Razor: a low-power pipeline based on circuit-level timing speculation. In: IEEE/ACM International Symposium on Microarchitecture, pp. 7–18 (2003)
9. Das, S., Tokunaga, C., Pant, S., Ma, W.-H., Kalaiselvan, S., Lai, K., Bull, D.M., Blaauw, D.T.: RazorII: In situ error detection and correction for PVT and SER tolerance. *IEEE J. Solid State Circuits* **44**(1), 32–48 (2009)
10. T. Nakura, K. Nose, M. Mizuno, Fine-grain redundant logic using defect-prediction flip-flops. In: IEEE International Solid-State Circuits Conference, pp. 21–23 (2007)
11. Bowman, K.A., Tschanz, J.W., Kim, N.S., Lee, J.C., Wilkerson, C.B., Lu, S.L.L., Karnik, T., De, V.K.: Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance. *IEEE J. Solid State Circuits* **44**(1), 49–63 (2009)
12. Das, B.P., Onodera, H.: Frequency-independent warning detection sequential for dynamic voltage and frequency scaling in ASICs. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **22**(12), 2535–2548 (2014)
13. Firouzi, F., Ye, F., Chakrabarty, K., Tahoori, M.B.: Aging-and variation-aware delay monitoring using representative critical path selection. *ACM Trans. Des. Autom. Electron. Syst.* **20**(3), 39:1–39:23 (2015)
14. Chan, T.-B., Gupta, P., Kahng, A. B., Lai, L.: Synthesis and analysis of design-dependent ring oscillator (DDRO) performance monitors. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **22**(10), 2117–2130 (2014)

15. Park, J., Abraham, J.A.: A fast, accurate and simple critical path monitor for improving Energy-Delay product in DVS systems. In: IEEE/ACM International Symposium on Low Power Electronics and Design, pp. 391–396 (2011)
16. Drake, A., Senger, R., Deogun, H., Carpenter, G., Ghiasi, S., Nguyen, T., James, N., Floyd, M., Pokala, V.: A distributed critical-path timing monitor for a 65nm high-performance microprocessor. In: IEEE International Solid-State Circuits Conference pp. 398–399 (2007)
17. Tschanz, J., Bowman, K., Walstra, S., Agostinelli, M., Karnik, T., De, V.: Tunable replica circuits and adaptive voltage-frequency techniques for dynamic voltage, temperature, and aging variation tolerance. In: IEEE Symposium on VLSI Circuits, pp. 112–113 (2009)
18. Drake, A.J., Floyd, M.S., Willaman, R.L., Hathaway, D.J., Hernandez, J., Soja, C., Tiner, M.D., Carpenter, G.D., Senger, R.M.: Single-cycle, pulse-shaped critical path monitor in the POWER7+ microprocessor. In: Proceedings of the International Symposium on Low Power Electronics and Design pp. 193–198 (2013)
19. Tokunaga, C., Ryan, J.F., Karnik, T., Tschanz, J.W.: Resilient and adaptive circuits for voltage, temperature, and reliability guardband reduction. In: IEEE International Reliability Physics Symposium, pp. 1–5 (2014)
20. Drake, A.J., Senger, R.M., Singh, H., Carpenter, G.D., James, N.K.: Dynamic measurement of critical-path timing. In: IEEE International Conference on Integrated Circuit Design and Technology, pp. 249–252 (2008)
21. Ikenaga, Y., Nomura, M., Suenaga, S., Sonohara, H., Horikoshi, Y., Saito, T., Ohdaira, Y., Nishio, Y., Iwashita, T., Satou, M., Nishida, K., Nose, K., Noguchi, K., Hayashi, Y., Mizuno, M.: A 27% Active-Power-Reduced 40-nm CMOS multimedia SoC with adaptive voltage scaling using distributed universal delay lines. *IEEE J. Solid State Circuits* **47**(4), 832–840 (2012)
22. Islam, A.K.M.M., Tsuchiya, A., Kobayashi, K., Onodera, H.: Variation-sensitive monitor circuits for estimation of global process parameter variation. *IEEE Trans. Semicond. Manuf.* **25**(4), 571–580 (2012)
23. Fujimoto, S., Mahfuzul Islam, A.K., Matsumoto, T., Onodera, H.: Inhomogeneous ring oscillator for within-die variability and RTN characterization. *IEEE Trans. Semicond. Manuf.* **26**(3), 296–305 (2013)
24. Iizuka, T., Asada, K.: All-digital PMOS and NMOS process variability monitor utilizing shared buffer ring and ring oscillator. *IEICE Trans. Electron.* **E95-C**(4), 627–634 (2012)
25. Kim, T.T.H., Lu, P.F., Jenkins, K.A., Kim, C.H.: A ring-oscillator-based reliability monitor for isolated measurement of NBTI and PBTI in high-k/metal gate technology. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **23**(7), 1360–1364 (2015)
26. Nishizawa, S., Onodera, H.: A ring oscillator with calibration circuit for on-chip measurement of static IR-drop. *IEEE Trans. Semicond. Manuf.* **26**(3), 306–313 (2013)
27. Krishnaswamy, V., Brooks, J., Konstadinidis, G., Curtis, M., Pham, H., Turullols, S., Shin, J., Yifan, Y., Zhang, H.: Fine-grained adaptive power management of the SPARC m7 processor. In: IEEE International Solid-State Circuits Conference, pp. 1–3 (2015)
28. Mair, H.T., Gammie, G., Wang, A., Lagerquist, R., Chung, C., Gururajao, S., Kao, P., Rajagopalan, A., Saha, A., Jain, A., et al.: A 20nm 2.5 GHz ultra-low-power tri-cluster CPU subsystem with adaptive power allocation for optimal mobile SoC performance. In: IEEE International Solid-State Circuits Conference, pp. 76–77 (2016)
29. Hokimoto, S., Ishihara, T., Onodera, H.: Minimum energy point tracking using combined dynamic voltage scaling and adaptive body biasing. In: IEEE International System-on-Chip Conference, pp. 1–6 (2016)
30. Islam, A.K.M.M., Onodera, H.: Area-efficient reconfigurable ring oscillator for device and circuit level characterization of static and dynamic variations. *Jpn. J. Appl. Phys.* **53**(4S), 04EE08 (2014)
31. Islam, A.K.M.M., Shiomi, J., Ishihara, T., Onodera, H.: Wide-supply-range all-digital leakage variation sensor for on-chip process and temperature monitoring. *IEEE J. Solid State Circuits* **50**(11), 2475–2490 (2015)

32. Bowman, K., Tschanz, J., Lu, S., Aseron, P., Khellah, M., Raychowdhury, A., Geuskens, B., Tokunaga, C., Wilkerson, C., Karnik, T., et al.: A 45 nm resilient microprocessor core for dynamic variation tolerance. *IEEE J. Solid State Circuits* **46**(1), 194–208 (2011)
33. Kim, Y., Shin, D., Lee, J., Lee, Y., Yoo, H.J.: A 0.55 V 1.1 mW artificial intelligence processor with on-chip PVT compensation for autonomous mobile robots. *IEEE Trans. Circuits Syst. Regul. Pap.* **65**(2), 567–580 (2018)
34. Mahfuzul, I., Kamae, N., Ishihara, T., Onodera, H.: A built-in self-adjustment scheme with adaptive body bias using P/N-sensitive digital monitor circuits. In: *IEEE Asian Solid State Circuits Conference*. 101–104 (2012)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.