An improved timing error prediction monitor for wide adaptive frequency scaling

Weiwei Shan\textsuperscript{1}, Xinning Liu\textsuperscript{1}, Minyi Lu\textsuperscript{1}, Shuai Shao\textsuperscript{1}, Zhikuang Cai\textsuperscript{2}, and Jun Yang\textsuperscript{a)}

\textsuperscript{1} National ASIC System Engineering Research Center, Southeast University, Nanjing, China, 210096
\textsuperscript{2} College of Electronic Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, 210023
\textsuperscript{a)} dragon@seu.edu.cn

Abstract: In order to eliminate the excess timing margin in integrated circuit due to PVT variations, we propose a low overhead timing error prediction monitor, which has only 14 transistors with negligible power overhead. It can generate a predicted alarm signal when the timing is intense before real errors occur. A bunch of timing monitors are inserted at the end of selected critical paths. When there are timing prediction signals, the system clock will be stretched immediately to avoid real timing errors. Applied on a computation intensive Bitcoin Miner chip under 40 nm CMOS process, the simulation results show that it can operate at a wide voltage range of 0.5–1.1 V. This timing prediction monitor based adaptive frequency scaling system can increase the frequency to 2.1\times at near-Vth voltage and 1.21\times at super-Vth region compared to the original non-monitored circuit. Thus, it is an effective way to mitigate the effect of PVT variations.

Keywords: timing monitor, low power, wide voltage range

Classification: Integrated circuits

References

[1] D. Ernst, \textit{et al.}: “Razor: A low-power pipeline based on circuit-level timing speculation,” MICRO (2003) 7.
[2] M. Eireiner, \textit{et al.}: “In-situ delay characterization and local supply voltage adjustment for compensation of local parametric variations,” IEEE J. Solid-State Circuits \textbf{42} (2007) 1583 (DOI: 10.1109/JSSC.2007.896695).
[3] K. Seongjong, \textit{et al.}: “Razor-lite: A side-channel error-detection register for timing-margin recovery in 45 nm SOI CMOS,” ISSCC (2013) 264.
[4] S. Kim, \textit{et al.}: “Variation-tolerant, ultra-low-voltage microprocessor with a low-overhead, within-a-cycle in-situ timing-error detection and correction technique,” IEEE J. Solid-State Circuits \textbf{50} (2015) 1478 (DOI: 10.1109/JSSC.2015.2418713).
[5] K. A. Bowman, \textit{et al.}: “A 45 nm resilient microprocessor core for dynamic variation tolerance,” IEEE J. Solid-State Circuits \textbf{46} (2011) 194 (DOI: 10.1109/JSSC.2010.2089657).
[6] Z. Hao, \textit{et al.}: “Ultra low overhead error mask flip-flop in near threshold voltage processor design,” Electron. Lett. \textbf{52} (2016) 1799 (DOI: 10.1049/
1 Introduction

Process, voltage, and temperature (PVT) variations have been an important issue in integrated circuit design, which may cause timing errors in worst cases. To prevent potential timing failure due to PVT variations, sufficient timing/voltage margin is usually added in IC design-time, which leads to waste of performance and power consumption. In order to reduce the timing margin, two kinds of in-situ timing monitoring techniques have been proposed [1, 2, 3, 4, 5, 6, 7, 8, 9]. One is timing error detection and correction (EDAC) [1, 3, 4, 5], represented by Razor series [1, 3], which can detect timing errors in critical paths and correct them later. The other is timing error prediction [2, 7, 8], such as Canary Flip-Flop [2], which can predict the occurrence of timing errors just before real errors occur. However, Razor [1] and Canary Flip-Flop [2] needed at least 28–44 transistors for timing error monitoring, which introduced a lot of area overhead. Razor-lite [3] was proposed with only 8 additional transistors, but Razor-lite could not work properly at the near-threshold voltage due to its threshold loss. Sparse error detection latch (EDL) [4] optimized the structure of Razor-lite to make it be able to work at 0.5 V, but it introduced a large overhead (24 transistors). Plus, it was not for wide operating range. Since the best energy efficiency operation for most digital circuits is at the near-threshold voltage region [4], but the high performance locates at super-threshold region, a timing monitor supports wide-operating-range application becomes important. Seldom research focused on this wide operating range timing monitoring technique. In order to design a timing monitor suitable for wide voltage operation with fewer transistors, we propose a compact timing prediction monitor with only 14 additional transistors. It has no need of additional error correction mechanism, thus its overall area overhead is small. The proposed timing monitor is implemented on a Bitcoin Miner chip under SMIC 40 nm CMOS process. Simulation results show that it can operate at a wide voltage range of 0.5–1.1 V, with effective frequency gain compared to the non-monitoring system.

2 Circuit design

The function of the prediction timing monitor is to detect whether the input data
arrives during the “detection window”. Here the detection window is usually a short period before the rising edge of the clock. If the input data changes in the detection window, it will be regarded as a potential timing error. In this case, the timing monitor will generate an alarm signal of ‘Pre.error’. It is usually placed in parallel of an endpoint FF in a critical path, as shown in Fig. 1(a). Its input is ‘Din’ and its output is the timing alarm signal of ‘Pre.error’.

The proposed circuit structure is shown in Fig. 1(a), which is mainly composed of an inverter with a transmission gate inserted in its middle node and a XOR gate. Its main working principle is: during the detection window, the transmission gate is shut off, then if an input data transition occurs, there will be a voltage difference between the two internal nodes of A and B. Thus, the XOR will output a ‘Pre.error’ signal. It includes 14 transistors in total, consisting of 4 CMOS transistors and a XOR (a typical 10-transistor structure in standard cell library).

The timing diagram of the monitor is illustrated as Fig. 1(b). The negative clock phase is used as the detection window. During the first clock period, the data Din arrives before the detection window which means the timing margin is enough. Since M3 and M4 are turned on in the positive clock phase, voltages at nodes A and B are the same, thus there is no timing alarm signal. During the second clock period, the data Din has a ‘1->0’ transition in the detection window, which means the timing is intense. Thus, node A is charged to ‘1’, but node B would remain ‘0’. Therefore, the XOR gate generates a ‘Pre.error’ signal. During the third clock period, a ‘0->1’ transition of ‘Din’ occurs in the negative clock phase, and the timing monitor generates a timing alarm signal too.

3 The timing monitor simulation results

This proposed structure has a potential risk at low voltage, that is, M1~M4 might be turned on simultaneously if the input data has a large transition time, which may cause a faulty alarm signal. Thus, the sizes of transistors are reconsidered to avoid this risk, as shown in Fig. 1(a), by using large sized M1 and M2. Its functionality and robustness are verified by 10k Monte Carlo simulations at 0.5 V and 0°C. As shown in Fig. 2(b), the proposed timing monitor can work reliably at 0.5 V.
We design the layout of the timing monitor as shown in Fig. 2(a), with an area of 2.139 µm², which is 0.47\( \times \) of a FF in the standard cell library. Since timing monitors are usually inserted in some selected critical paths in parallel with the endpoint FF, we compare a typical flip-flop in the standard library and our timing monitor combined with a flip-flop in Table I. The data come from HSPICE simulations under 0.5 V, SS corner, 0°C at 80 MHz frequency. It can be seen that the power overhead of our timing monitor combined with a FF is negligible (1.08× switch energy and 1.04× leakage power). Here the switch energy data are evaluated assuming 10% data switching rate as [3, 4].

|                  | Flip-flop in lib | Flip flop+timing monitor |
|------------------|------------------|--------------------------|
| Switch Energy    | 1.321 fJ         | 1.427 fJ (1.08×)         |
| Leakage Power    | 6.504 nW         | 6.769 nW (1.04×)         |
| Transistor       | 28               | 28+14                    |
| Area (µm²)       | 4.486 µm²        | 6.625 µm² (1.47×)        |

### 4 Circuit implementation and simulation

We implement the proposed timing monitor on a Bitcoin Miner chip under SMIC 40 nm CMOS process. Bitcoin Miner chip is a highly computation intensive circuit, which is mainly composed of parallel SHA-256 blocks and a controller. There is no SRAM in this chip. It has 2310 FFs and a core area of 0.21 mm².

As the system diagram shown in Fig. 3, a series of timing monitors are inserted at the end of selected critical paths. Then all the ‘Pre.error’ signals are ORed together by dynamic ORs to get a total error signal, which is sent to frequency control FSM to control the adaptive clock stretching block [10] to generate a stretched clock in order to avoid the real timing error. The Frequency control FSM also adjusts the duty cycle of the clock to set an appropriate monitoring window. In this chip, 211 FFs in the top 10% critical paths are inserted with our monitor among all 2310 FFs. The total area overhead is only 4.3% including frequency control module.

First, the function of the Bitcoin system is verified by digital design tools. However, they could not provide accurate monitoring based AFS simulation. Thus HSIM-VCS co-simulation is used to verify the AFS monitoring effect. Here the critical paths inserted with timing monitors and the adaptive clock stretching block
are simulated by HSIM, which can provide a transistor-level simulation with high accuracy. The other parts are simulated by VCS and interface with HSIM by providing the needed signal stimulus.

Fig. 4 is the whole timing prediction and AFS tuning process at super-Vth region. At first, the frequency (represented by ‘Freq_show_id’) gradually increases from 960MHz to 969MHz, where the first error prediction appears, as shown in Fig. 4(a). Then the ‘Pre_error_all’ signal activates the clock stretch procedure to make the clock cycle stretch immediately and last for a few cycles, as shown in Fig. 4(b). Thus, a potential timing error is avoided. The tuning process at near-Vth is similar.

In order to evaluate the performance benefit due to our AFS system, we first obtain the frequency baselines by simulating the system at the worst corner, the worst temperature and a 10% reduced supply voltage according to [3, 4]. The baselines are 88 MHz for near-Vth (0.6 V as the standard voltage) and 819 MHz for super-Vth (1.1 V as the standard voltage). We obtain the first warning frequency generated by the timing monitor at different corners and temperatures. As shown in Fig. 5(a) and Fig. 5(b), our timing monitor can obtain up to 2.1× frequency benefit.
at near-Vth and obtain up to 1.21× frequency benefit at super-Vth. Because the circuit suffers more PVT variations at near threshold voltage, this timing prediction based AFS technique can obtain more performance benefits at near-Vth region.

The comparisons of other timing monitors are listed in the up-half part of Table II. Our proposed timing monitor can operate at a wide voltage range of from 0.5 V ~1.1 V, with a relative small number of transistors and little power overhead. Plus, it does not need to change the original flip-flop. Thus, it has fewer impacts on the original circuit, leading to a simplified design flow and a small area overhead.

Table II. Comparisons of our timing monitor and AFS system with other works

|                  | DSTB [5] | Razor-lite [3] | Sparse EDL [4] | Proposed |
|------------------|----------|----------------|---------------|----------|
| Timing monitor   |          |                |               |          |
| Transistors      | 26       | 8              | 24            | 14       |
| Switch power     | 14%~34%  | 2.7%           | \             | 8%       |
| Voltage range    | 0.6 V~1.1 V | 0.78 V~1.1 V  | 0.5 V        | 0.5 V~1.1 V |
| Change Flip-Flop | No       | Yes            | No            | No       |
| Process          | 45 nm    | 45 nm          | 65 mm         | 40 nm    |
| Insertion rate   | 12%      | 19.8%          | 13%           | 9.1%     |
| Performance gain | 41%      | /              | 130%          | 21% to 110% |
| Area overhead    | 3.8%     | 4.42%          | 7.8%          | 4.3%     |

5 Conclusion

In this letter, we propose a timing monitor with only 14 transistors which can work with a wide voltage range. The timing monitor was implemented in a Bitcoin Miner chip and the system was evaluated in SMIC 40 nm CMOS process with only 4.3% area overhead. At 0.55 V, the proposed timing monitor system can operate at up to 2.1× frequency. Therefore, it is an effective way to decrease the timing margin in digital IC design caused by PVT variations.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (61574033, 61774038 and 61504065), the Natural Science Foundation of Jiangsu Province (BK20150848), National 863 project of China (2015AA016601) and the National Science and Technology Major Project (2014ZX01030-101).