An Adaptive Body-Biasing SoC using in situ Slack Monitoring for Runtime Replica Calibration

Mehdi Saligane1, Jeongsup Lee1, Qing Dong2, Makoto Yasuda3, Kazuyuki Kumen3, Fumitaka Ohno3, Satoru Miyoshi3, Masaru Kawamini1,2,3, David Blaauw4, Dennis Sylvester1

1University of Michigan, Ann Arbor, MI 2Mie Fujitsu Semiconductor Limited 3Fujitsu Electronics America, Inc.

Abstract

This work proposes a hybrid approach combining the benefits of in situ timing slack monitoring and tunable replica techniques while avoiding their drawbacks (e.g., mistracking, high overhead). In a 55nm technology with strong body bias coefficient, we demonstrate an adaptive Cortex-M0 that is based on an in situ assisted tunable replica circuit and shows tracking error of <2% across [0.5–0.9]V supply and [−40–125]°C, achieving up to 53% energy improvement.

Introduction & Approach Overview

To address rising power densities in scaled processes, recent designs have explored near-threshold (NT) operation, in which supply voltage is aggressively scaled to increase energy efficiency [2,8]. However, NT operation exacerbates sensitivities to process, voltage and temperature (PVT) variations. As a result, designs often use large margins in frequency and power to meet functionality in worst-case conditions, wasting power/performance in typical scenarios.

One approach to reducing margins is the use of tunable replica circuits (TRC) as a proxy for SoC critical path delay, which is difficult to measure during operation [1-5]. However, TRCs mistrack actual circuit behavior across voltage and temperature (VT), limiting gains. Alternatively, Razor-style approaches [7] use in situ delay monitoring that removes all PVT margins. However, such “let fail and correct” approaches incur large area, power, and/or design complexity overheads. To enable tight performance tracking across PVT variations while limiting overhead, we propose a new adaptive design approach that dynamically calibrates TRCs at runtime using in situ transition detecting flip-flops (TD-FFs) and a novel slack merging cell that instrument the SoC design. Each time a new VT operating condition is encountered (based on temperature and voltage sensors), the SoC is briefly halted, allowing pre-stored worst-case test patterns to be executed while critical path delays are monitored by TD-Flipflops. The observed delay slack is used to calibrate the TRCs, after which the SoC resumes regular operation (flow diagram in Fig. 2). TRCs (along with p-well/n-well charge-pump) are used in a control loop to compensate operating condition VT fluctuations while maintaining constant speed. Despite large CPU logic depth, wire delay causes significant TRC mistracking across wide range VT operation. Hence the TD-FFs instrument two distinct groups; logic and interconnect (RC) delay-dominated (Fig. 1). We experimentally demonstrate that these two sets of paths have significantly different response to PVT variations and thus by separately observing their delay with TD-FFs we can more accurately calibrate the TRCs and reduce margin. The joint use of tight cycle-to-cycle delay mistracking and pre-stored worst-case test patterns allows energy savings up to 53% (Table I) while maintaining safety operation at extreme conditions (e.g., 0.5V, [−40, 125]°C).

Proposed Circuits and Implementation

Fig. 1 shows the organization of the proposed adaptive SoC while Fig. 2 shows the associated runtime TRC calibration process. The proposed approach is applied to a Cortex-M0 processor and AHB-lite system with 32KB RAM and 32KB ROM in 55nm Deeply Depleted Channel (DDC) CMOS. The design operates from 0.5 – 0.9V and [−40–125]°C, achieving up to 53% energy improvement.

The aforementioned calibration comparator compares the worst-case slack from the TD-FFs to that from the TRCs, sending an error signal to the control architecture if slack is insufficient. The calibration controller initiates execution of the worst-case test pattern vectors on the SoC and then performs a binary search for TRC weights that create matched slack between TD-FFs and TRCs (see Fig. 3). The logic and RC dominated delay groups are independently merged by a TD-merge cell. During regular operation, the adaptive-BB controller monitors the TRCs to compare Vslack_trc against two preset thresholds (Vwarning and Verror). If Vslack_trc is lower than Verror, a low-slew flag triggers forward body biasing of the n-well/p-well through the integrated charge-pump (see Fig. 4). Similarly, when Vslack_trc exceeds both values, reverse body bias is slowly incremented. Finally, if Vslack_trc lies between Vwarning/Verror, no action is taken.

Measurement Results

Fig. 5 highlights the two sets of in situ monitored paths determined at design time; logic dominant paths are critical at 0.5V while RC dominant paths are critical at 0.9V (top). The measured error count (Fig. 5 bottom; here Vslack tracks actual timing slack) generated by the two sets of TD-FFs emphasizes the need to instrument both types of paths even though at a particular voltage, only one type might limit performance. The results also indicate that the TRC must transition from RC to logic weightings as operating voltage changes. Fig. 6 shows TRC tracking to measured core Fmax when calibrated at 0.5V and 0.9V at 25°C or alternatively at 0.9V for −40°C and 125°C, following the procedure in [1]. This calibration incurs mistracking of 12–34% in worst-case conditions (−40°C, 0.5V). In contrast, the proposed auto-tuned TRC shows <2% mistracking across 0.5–0.9V and −40°C to 125°C. Fig. 7 shows the Cortex-M0 shmmu plot (300MHz at 0.9V) and the measured energy gain when operating with the proposed self-tuned TRC vs. a TRC calibration at 0.5 and 0.6V, −40°C and 125°C operation. Finally, Fig. 8 shows measured results of the full closed-loop body biased system that automatically responds to voltage/temperature fluctuations. The proposed adaptive-BB approach with self-tuned TRCs shows high fidelity in tracking and mitigating PVT variations, resulting in up to 53% energy savings. The approach is most beneficial at low voltage/temperature at which the baseline core has poor performance and becomes a bottleneck for energy efficiency in conventional NT systems.

References
