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

Mehdi Saligane[1], Jeongsup Lee[1], Qing Dong[1], Makoto Yasuda[2], Kazuyuki Kumeno[2], Fumitaka Ohno[2], Satoru Miyoshi[3], Masaru Kawaminami[2,3], David Blaauw[1], Dennis Sylvester[1]

[1]University of Michigan, Ann Arbor, MI [2]Mie Fujitsu Semiconductor Limited [3]Fujitsu 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-flip-flops. 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 over the PVT range, together with closed-loop body biasing allows energy savings up to 53% (Table I) while maintaining safe 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 exploits the strong body bias coefficient of the target technology to re-center the V_{th} shift due to PVT variations shown in Fig. 1 (top). The proposed approach has low overhead since the number of TD-FFs is low (100s of FFs) and in typical operation their delay monitoring is disabled, avoiding power overhead. No TRC pre-calibration is required since the SoC is fully self-calibrating, generating TRC weights at every new condition, including at boot time thereby greatly reducing testing costs. Further, since TRC tuning weights are stored for all previously encountered PVT conditions, processor halting to perform calibration and run worst-case vectors becomes increasingly infrequent over time, reducing performance overhead.

A key design element of the proposed adaptive approach are the TD-FFs; they produce pulses of a duration that is proportional to the observed slack at their inputs. These pulses are fed to a "slack to V combiner" (labeled TD-merge cell in figures) that transmits the worst-case slack to a calibration comparator in the form of a voltage pulse. This pulse is transformed to V_{slack} by charging a tunable capacitance. Measured results in Fig. 4 (bottom right) show the relationship between timing slack and V_{slack}. As seen in Fig. 4 the TD-FF consists of a simple XOR gate that compares the incoming data with the already latched data. If data is transitioning, a short pulse is generated until the flip-flop latches the new data. In this way, pulse length becomes proportional to slack of the data transition. The TD-merge cell functions properly in simulation down to 300mV. Pulse generation is disabled when the TD-FF clock is gated by clock gating cell CG. TD-FF pulses are combined with an AND gate (2-level in the test chip) to produce a single pulse corresponding to the shortest slack across fan-in TD-FFs.

The aforementioned calibration comparator compares the worst-case slack from the TD-FFs to that from the TRCs, sending an error signal to the calibration controller 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 V_{slack_trc} against two preset thresholds ($V_{warning}$ and V_{error}). If V_{slack_trc} is lower than V_{error}, a low-slack flag triggers forward body biasing of the n-well/p-well through the integrated charge-pump (see Fig. 4). Similarly, when $V_{slack\text{-}trc}$ exceeds both values, reverse body bias is slowly incremented. Finally, if V_{slack_trc} lies between $V_{warning}/V_{error}$, 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 V_{slack} 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 F_{MAX} 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 shmoo 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

[1] M. Cho *et al*. ISSCC, 2016
[2] M. Nomura *et al*, VLSI, 2013
[3] S. Clerc *et al*, ISSCC, 2015
[4] R. Wilson *et al*. ISSCC, 2014
[5] K. Wilcox *et al*, ISSCC, 2015
[6] M. S.Floyd *et al*, ISSCC, 2017
[7] Y. Zhang *et al*, ISSCC, 2016
[8] J. Myers *et al*, VLSI, 2017

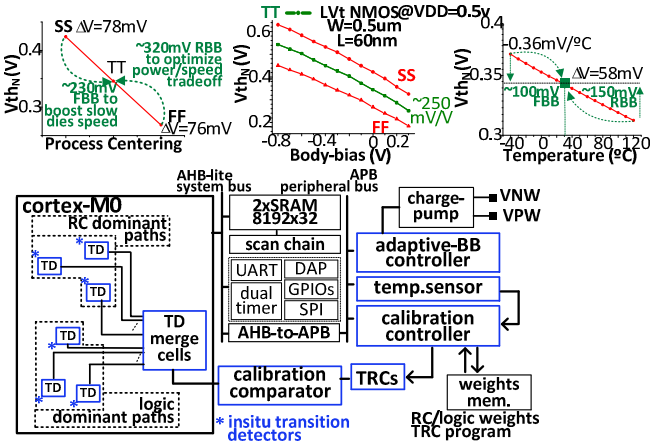

Figure 1. Vth design centering is shown using body-biasing to compensate temperature or process fluctuations. System overview.

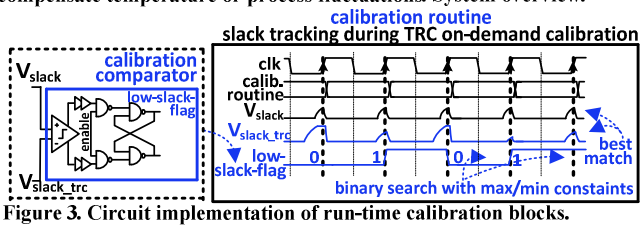

Figure 3. Circuit implementation of run-time calibration blocks.

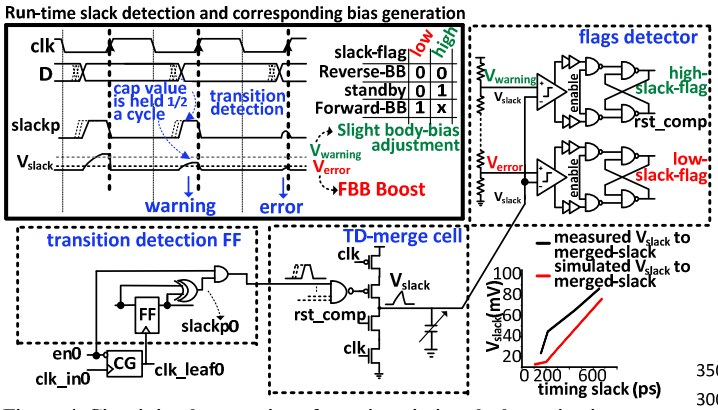

Figure 4. Circuit implementation of run-time timing slack monitoring.

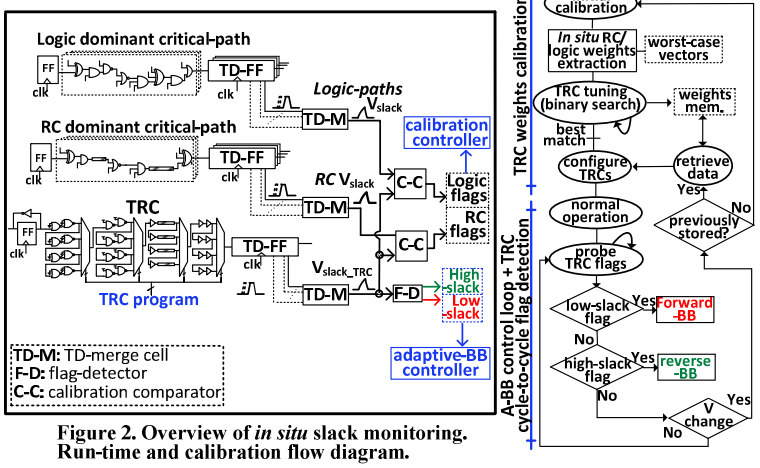

Figure 2. Overview of *in situ* slack monitoring. Run-time and calibration flow diagram.

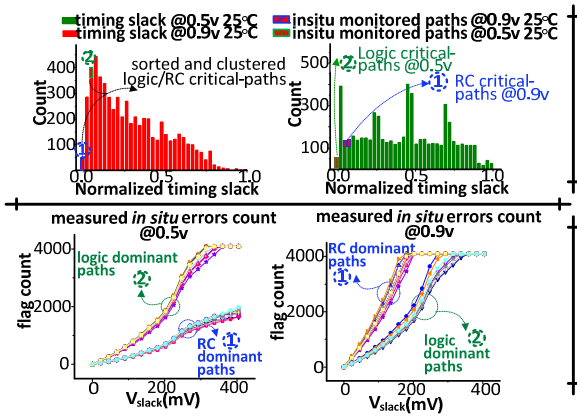

Figure 5. Slack histogram at design sign-off (top). Timing slack measured versus RC/logic flag counts using *in situ* slack detection at 0.5V or 0.9V operating supply voltages (bottom).

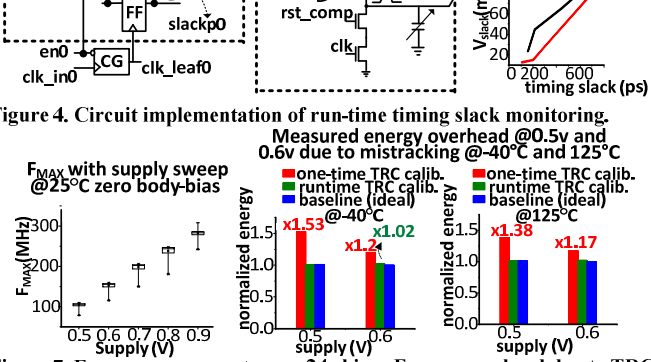

Figure 7. Fmax measurements over 24 chips. Energy overhead due to TRC mistracking using one-time, 1-point calibration at -40°C and 125°C, at 0.5V and 0.6V supply voltage versus mistracking with runtime TRC calibration.

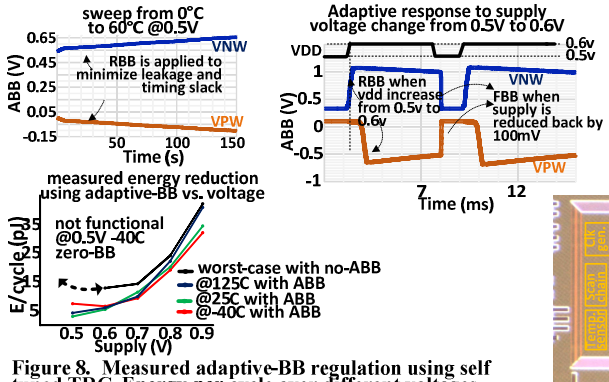

Figure 8. Measured adaptive-BB regulation using self tuned TRC. Energy-per-cycle over different voltages and temperature conditions with and w/o adaptive-BB.

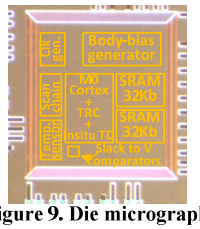

Figure 9. Die micrograph.

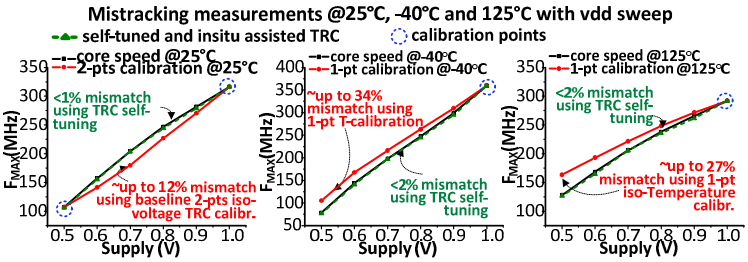

Figure 6. Measured core frequency versus supply, illustrating TRC frequency mistracking at different conditions (V,T) when using: one-point TRC calibration, using multi-points and proposed self-tuned TRC using *in situ* slack detection.

Table I. Comparison Table

Parameters		[1] ISSCC16	[2] VLSI13	[3] ISSCC15	[4] ISSCC15	[8] VLSI17	This work
Process (nm)		22	40	28 FDSOI	28 FDSOI	65	**55 DDC**
Technique		Tunable Replica	Replica ROSC+CPR	Tunable Replica	TDC+ Tunable Replica	Tunable ROSC	**In situ+ Tunable Replica**
Detection		Pre-edge	Pre-edge	Pre-edge	Pre-edge	correlation	**Pre-edge**
Host CPU		Graphics exec. core	32b CPU+ SIMD	SPARC V8	DSP+MAC	ARM cortex-M0+	**ARM cortex-M0**
SoC Area (mm²)		3.38	25	0.62	1	3.76	**1.32**
Closed-loop type		External AVS	AVS+AFS	ABB 0 to ± 1.4	-	DVFS	**ABB 0.3 to -0.8**
Supply (V)		0.4~0.9	0.45	0.33~0.45	0.397~1.3	0.3~0.8	**0.5~0.9**
Frequency (MHz)		800	7.5	1~20	460~2600	0.012~60	**106~315**
E/cycle (pJ)	core	-	-	-	-	-	**0.83**
	+periph.	-	568	-	62	6.3	**5.2**
	+IVR	-	-	26	-	12.9	**-**
Calibration Per die		2-pt	-	Multi-pt	Multi-pt	-	**On-chip self-tuning**
Mismatch (%)	1-pt calib.	5% of V_{dd}	-	1.7 @calib. pt	8~4 @calib. pt	5.3 @calib. pt	**<2%@125C 34%@-40C**
	2-pt calib.	2.5% of V_{dd}	-	-	-	-	**12%@25°C**
	Proposed self-tuning	-	-	-	-	-	**<2% from [-40~125]°C**
Energy savings		4%@0.8V 14%@0.4V	6% (33%Fmax)	58%	40.6%	24%	**38%@125C 53%@-40C**