9.8 In Situ Delay-Slack Monitor for High-Performance Processors Using An All-Digital Self-Calibrating 5ps Resolution Time-to-Digital Converter

David Fick, Nurrachman Liu, Zhiyoong Foo, Matthew Fojtik, Jae-sun Seo, Dennis Sylvester, David Blaauw

University of Michigan, Ann Arbor, MI

Advanced CMOS technologies have become highly susceptible to process, voltage, and temperature (PVT) variation. The standard approach for addressing this issue is to increase timing margin at the expense of power and performance. One approach to reclaim these losses relies on canary circuits [1] or sensors [2], which are simple to implement but cannot account for local variations. A more recent approach, called Razor, uses delay speculation coupled with error detection and correction to remove all margins but also imposes significant design complexity [3]. In this paper, we present a minimally-invasive *in situ* delay slack monitor that directly measures the timing margins on critical timing signals, allowing margins due to both global and local PVT variations to be removed.

At the heart of the monitor is an all-digital, self-calibrating time-to-digital converter (TDC) consisting of a 30-bit Vernier chain (VC) with each stage tuned to 5ps increments. The overall measurement window is 150ps, which is sufficient for timing slack measurements in 2 Ghz+ processor systems. A tunable delay chain is also included to allow a wider range of measurements, if necessary. A key concern in TDCs is the need for time consuming and complex calibration, which is especially difficult given the very high resolution demanded by high-performance processors. To avoid expensive off-chip measurements and tester procedures, we propose a new approach that allows the TDC to be automatically self-calibrated under the sole control of the processor using only the off-chip crystal oscillator. We use a process of statistical sampling to convert the relatively slow system clock to 1ps accurate calibration steps. The proposed approach was implemented on a 64-bit Alpha processor and can complete full self-calibration of the entire TDC in five minutes with approximately 1ps accuracy. Calibration can be performed in the field during idle periods, incurring no additional tester time or system downtime. Typical delay slack margins remain in place until the TDC is calibrated.

The TDC architecture is shown in Fig. 9.8.1. A measurement is initiated using the *start* signal, which allows the next clock edge to pass through a latch and into the device. The clock passes through the tunable delay chain, which allows coarse alignment of the detection window. The VC is calibrated so that each stage delays the clock by 5ps relative to the data, resulting in a forward sampling of time. The clock then exits the VC and triggers a *TDC done* signal, which requests that the processor read the latch states. Each delay element in the VC is tunable using eight identical processor-controlled capacitor loads, which are designed to induce 1ps shifts in delay. A uniform load structure was chosen in lieu of a more compact binary weighted scheme to ensure monotonicity, which greatly eases automation of calibration.

Narrow pulses (glitches) can occur on data signals and must be monitored since they could be captured by a pipeline register if delay slack margins are reduced too far. To prevent pulses from collapsing in the VC due to rise-time/fall-time imbalance, the chain lengthens pulses by accelerating falling transitions forward using *speedup* signals as shown in Fig. 9.8.1. Because of this, the TDC is calibrated for rising transitions - we include an XOR gate in order to monitor falling transitions. The metric of interest is the slack between the latest arriving input signal and the clock edge at the flip-flop (FF) input. To compute this metric, a mux local to the monitored FF selects between clock and data, and we subtract the two measured delay values. The number of FFs that need to be monitored depends on the balance of the pipeline; typically 30% of all FFs is sufficient [3].

The VC is calibrated using a reference data transition (*ref_tran*) generated by the reference delay chain (RDC in Fig. 9.8.1). The RDC is trigged by the clock and feeds test transitions into the data input of the VC. First, we find the zero point of the RDC (which is tunable in 1ps increments) by increasing delay until the zeroth bit of the VC shows equal probability of reading zero and one. The goal is then to shift *ref_tran* in exact increments of 5ps and tune the subsequent bits of the VC in the same manner. To accurately measure shifts in *ref_tran* under PVT variation, a measureable reference pulse (*ref_pulse*) is generated by subtracting *clock* and *ref_tran*. The signal *ref_pulse* is intentionally lengthened to ensure non-zero pulses when *ref_tran* and *clock* are aligned. As explained below, the length of *ref_pulse* is computed by comparing it to a known calibration pulse (*cal_pulse*) using statistical sampling, which is generated using a reference off-chip crystal. Although the exact relationship between the length of *ref_pulse* and the alignment of *ref_tran* cannot be accurately determined, it is unimportant since we require only that a 5ps shift in *ref_tran* will result in an identical 5ps shift of the *ref_pulse* length. Hence, we search for an RDC setting that produces an exact 5ps increase in *ref_pulse*. This ensures that *ref_tran* is delayed by an exact 5ps step and allows the next bit of the VC to be tuned.

To determine the length of *ref_pulse*, a start signal generates a single *ref_pulse* and *cal_pulse*, which are each fed to the enable input of identical up counters. A slow asynchronous sampling clock generated locally provides the clock input for both counters. Since it has no relationship to either pulse source, it statistically samples both signals over many trials (1M in our tests). This technique allows us to compare the length of *ref_pulse* relative to *cal_pulse* with sub-1ps accuracy despite pulse widths of hundreds of ps. Calibration error is not cumulative as the process moves through the VC since each *ref_pulse* is compared to the length of the zeroth-bit pulse.

Figure 9.8.7 shows the implementation of the proposed slack monitor in an Alpha processor fabricated in a 45nm process. The core interacts with the TDCs through memory mapped IO. In Fig. 9.8.3 we analyze the accuracy of the RDC statistical sampling method. After setting the RDC to a fixed phase pulse width (min, middle, max), we swept a range of calibration counter values and plotted the standard deviation of the reference counter after the calibration counter reaches its target value; trials continued until the calibration counter reached a particular value and then stopped, resulting in a corresponding reference counter value. The stability of readings indicates a deviation below 1ps for a calibration counter of 500,000 samples. Additional accuracy can be gained with exponentially larger calibration counter values, or by using stronger statistical methods than averaging. In this implementation, the reference clock was implemented with a current starved ring oscillator; a crystal oscillator would provide further accuracy.

In Fig. 9.8.4 we show the calibration settings of a VC delay chain. The "skew" signal for the clock chain side is enabled to provide a baseline delay of 5ps. To address local variation, the 1ps calibration capacitors are enabled as needed on either chain to obtain equal probability of a zero or one reading when the reference delay chain is stepped with 5ps increments. In this experiment we ran 10,000 trials to check the 50% probability condition (corresponding to a count of 5000) for each bit in the VC chain. Each time the condition is checked we try to find an improvement by enabling another capacitor on one side of the chain. For instance, if a count of 8000 is recorded, then the data transition is coming before the clock transition too frequently, so we enable a data capacitor to try to reduce the count. Once there are no more capacitors available, the configuration that yielded the best 50% condition is restored and the next bit of the VC is calibrated.

In Fig. 9.8.5 we use the reference delay chain to sweep 1ps step inputs through the VC and show output codes. Monotonicity is guaranteed due to the calibration method and also shows excellent linearity. Finally, the use of the TDC is illustrated by using it to measure a timing slack distribution of the most critical paths in the Alpha processor, as shown in Fig. 9.8.6, illustrating the potential use of the TDC.

Acknowledgements:
We thank ST Microelectronics for fabrication and ARM Ltd for their support and suggestions.

References:
[1] 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 65nm High-Performance Microprocessor," *International Solid-State Circuits Conference*, pp. 398-399, 2007.
[2] K. Woo, S. Meninger, T. Xanthopoulos, E. Crain, D. Ha, D. Ham, "Dual-DLL Based CMOS All-Digital Temperature Sensor for Microprocessor Thermal Monitoring," *International Solid-State Circuits Conference*, pp. 68-71, 2009.
[3] D. Blaauw, S. Kalaiselvan, K. Lai, W-H. Ma, S. Pant, C. Tokunaga, S. Das, D. Bull, "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance," *International Solid-State Circuits Conference*, pp. 400-401, 2008.

 978-1-4244-6034-2/10/$26.00

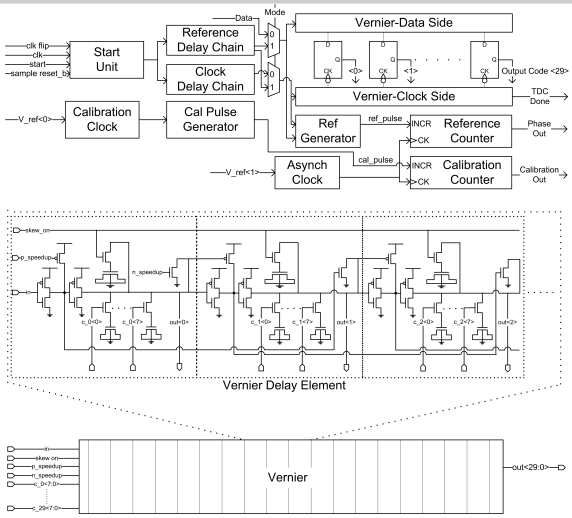


Figure 9.8.1: TDC architecture and Vernier delay chain implementation. The delay chain diagram is for a single side of the Vernier, and is also used in the Reference Delay Chain.

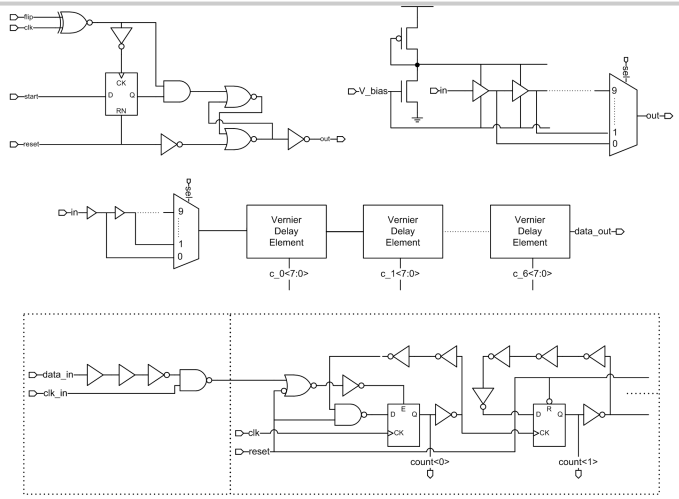


Figure 9.8.2: Start signal, clock delay chain, reference delay chain, reference pulse generator, and counter implementations. The first bit of the counter is decoupled from the remaineder of the counter to optimize for pulse capture speed.

9

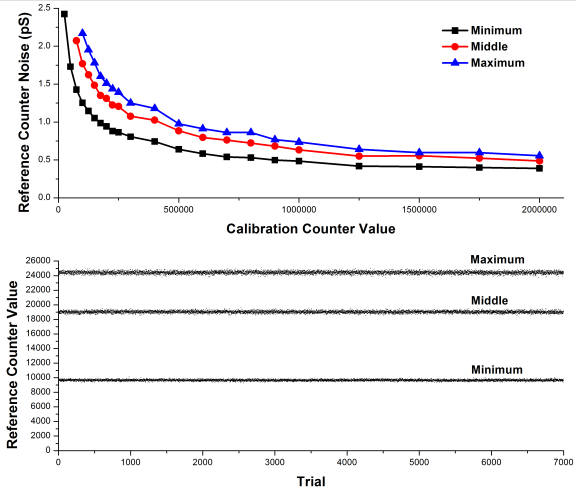


Figure 9.8.3: Reference counter stability measurements. For the minimum, maximum, and a middle fixed reference delay chain configuration, 7000 trials were run such that the calibration counter reached a value of 25,000 and the reference counter value was recorded.

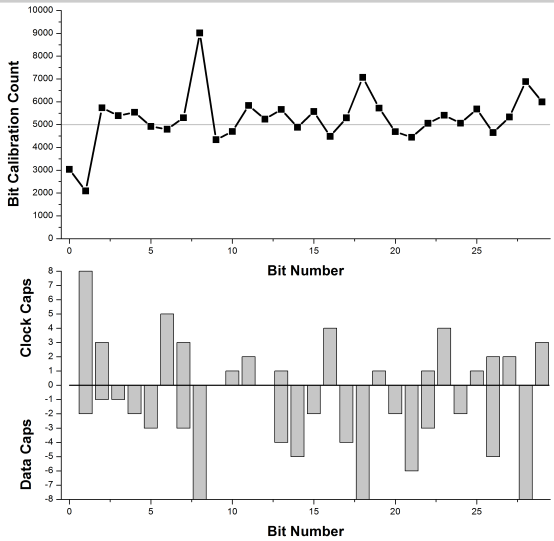


Figure 9.8.4: Calibration conversion counts and calibration settings with Vernier delay chain capacitors.

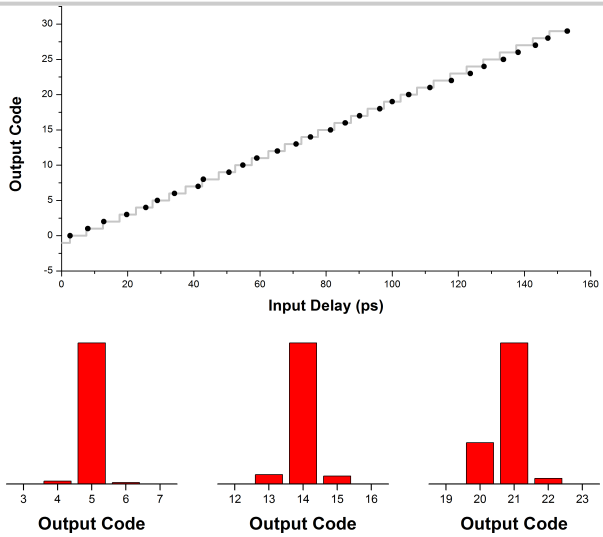


Figure 9.8.5: Input-output characteristics of the TDC and histogram of observed output codes illustrating expected noise level.

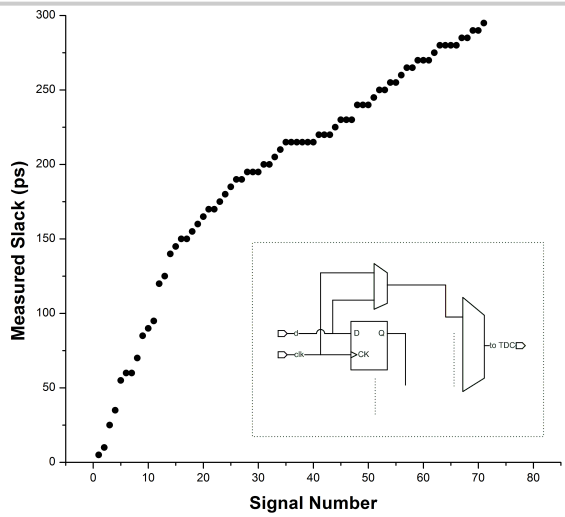
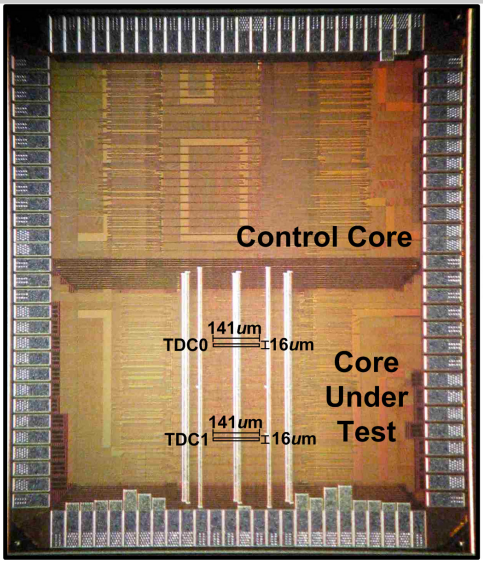


Figure 9.8.6: Measured slack distribution of paths relative to local clock. Circuit diagram shows path monitoring methodology, allowing the TDC to locally select between flip-flop clock and data.



Technology	45nm
TDC Area	141umx16um
Chip area	1.5mmx1.3mm

Figure 9.8.7: Die micrograph.

978-1-4244-6034-2/10/$26.00