

Compact Thermal-Diffusivity-Based Temperature Sensors in 40-nm CMOS for SoC Thermal Monitoring

Uğur Sönmez, *Member, IEEE*, Fabio Sebastiano, *Member, IEEE*, and Kofi A. A. Makinwa, *Fellow, IEEE*

Abstract—An array of temperature sensors based on the thermal diffusivity (TD) of bulk silicon has been realized in a standard 40-nm CMOS process. In each TD sensor, a highly digital voltage-controlled oscillator-based $\Sigma\Delta$ ADC digitizes the temperature-dependent phase shift of an electrothermal filter (ETF). A phase calibration scheme is used to cancel the ADC's phase offset. Two types of ETF were realized, one optimized for accuracy and one optimized for resolution. Sensors based on the accuracy-optimized ETF achieved a resolution of 0.36 °C (rms) at 1 kSa/s, and inaccuracies of ± 1.4 °C (3σ , uncalibrated) and ± 0.75 °C (3σ , room-temperature calibrated) from -40 °C to 125 °C. Sensors based on the resolution-optimized ETFs achieved an improved resolution of 0.21 °C (rms), and inaccuracies of ± 2.3 °C (3σ , uncalibrated) and ± 1.05 °C (3σ , room-temperature calibrated). The sensors draw 2.8 mA from supply voltages as low as 0.9 V, and occupy only 1650 μm^2 , making them some of the smallest smart temperature sensors reported to date, and well suited for thermal monitoring applications in systems-on-chip.

Index Terms—Phase-to-digital converter, temperature sensors, thermal diffusivity (TD), thermal monitoring, voltage-controlled oscillator (VCO)-based sigma-delta modulator.

I. INTRODUCTION

TODAY, microprocessors and other systems-on-chip (SoCs) employ billions of transistors that can switch at gigahertz rates. As a result, they can get hot enough to degrade their performance and even cause permanent damage. To avoid this, thermal management algorithms, driven by information from on-chip temperature sensors, slow them down or even shut them off when temperatures near reliability limits. To account for sensor errors, however, such algorithms must incorporate an appropriate safety margin. Given that the thermal resistance of a well-designed heat sink may be as low as 0.5 °C/W, a 5 °C margin corresponds to 10 W of unused power [1]. Since a typical microprocessor dissipates less than 100 W, this represents a significant loss of computing performance, and thus motivates the design of accurate temperature sensors. In multicore microprocessors, substantial thermal gradients and hotspots may occur, whose location is a dynamic function of workload. Thus, multiple on-chip temperature sensors are required, both to ensure reliability and to optimally spread the workload over different cores [2].

Manuscript received June 21, 2016; revised November 4, 2016; accepted December 20, 2016. Date of publication January 26, 2017; date of current version March 3, 2017. This paper was approved by Associate Editor Ken Suyama.

The authors are with the Electronic Instrumentation Laboratory/DIMES, Delft University of Technology, Delft 2628CD, The Netherlands.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSSC.2016.2646798

Since the location of hotspots cannot be easily predicted at design time, on-chip sensors must be small enough to be deployed in large numbers (up to 44 in modern microprocessors [3]), and for their position in the layout to be flexibly moved, even at a late stage of the development [2]. Accuracy requirements must be satisfied while minimizing the calibration effort, which could, otherwise, significantly increase manufacturing costs, especially when tens of sensors per chip are involved. The greatest accuracy is required around the reliability limit, with typical specifications being ± 1 °C at 70 °C, and only ± 3 °C at 50 °C [2]. In addition, to properly detect thermal transients with slopes as high as 0.5 °C/ms [2], sensor resolution must be significantly less than 0.5 °C, even with measurement times as short as 1 ms.

Most on-chip CMOS temperature sensors are currently based on parasitic p-n-p thanks to their relatively simple design and good energy efficiency. When implemented in nanometer CMOS, however, it has been shown that their inaccuracy is limited to only a few degrees Celsius, even after trimming [2], [4], [5]. Parasitic NPNs achieve better performance [6], [28], but are not available in baseline CMOS processes. Moreover, the base-emitter voltages of BJTs are about 0.7 V at room temperature, which makes it quite challenging to operate them from today's 1-V supplies. Other types of temperature sensors, e.g., based on resistors [7] or MOS transistors [8], [9], also exhibit poor inaccuracy when implemented in nanometer CMOS, and so must be combined with expensive multipoint temperature calibration.

As an alternative, the thermal diffusivity (TD) of bulk silicon can be used as a measure of temperature. This is strongly temperature-dependent (approximately proportional to $T^{-1.8}$) and well defined for the highly pure silicon used in ICs [10]. A TD-based temperature sensor (TD sensor) operates by measuring the time that it takes for heat pulses from a heater, usually a diffusion resistor, to diffuse through the substrate to a relative temperature sensor, usually a thermopile. This diffusion process can be modeled as an electrothermal low-pass filter, whose delay is in the order of a few microseconds for heater/thermopile spacings of a few micrometers. The corresponding phase shift is approximately proportional to absolute temperature ($\sim T^{0.9}$) [10].

The accuracy of an electrothermal filter (ETF) is mainly limited by variations in the spacing between the heater and the thermopile, which, in turn, is determined by the lithographic accuracy of the process used. Thus, the accuracy of TD sensors actually improves with technology scaling, as does the timing

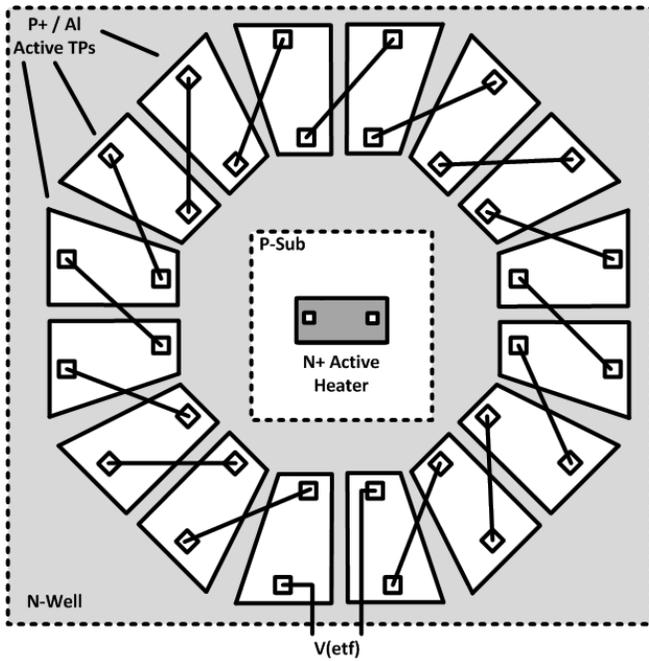


Fig. 1. Simplified layout of the proposed octagonal ETF in standard CMOS.

accuracy of their readout circuitry [11]. Moreover, since the required heat pulses can be generated from any supply voltage, TD sensors can be easily ported to newer technologies with lower supply voltages.

It has been shown that TD sensors can achieve untrimmed inaccuracy below $0.2\text{ }^{\circ}\text{C}$ in $0.18\text{-}\mu\text{m}$ CMOS [11]. However, the reported smart sensor was too large (0.18 mm^2) and too slow (1 Sa/s) for thermal monitoring applications. By employing more compact electronics, much smaller smart TD sensors with the areas of $8000\text{ }\mu\text{m}^2$ [12], and even $2800\text{ }\mu\text{m}^2$ [13], have been reported. However, these sensors were also implemented in a relatively mature $0.16\text{-}\mu\text{m}$ CMOS process.

This paper presents the first TD sensor realized in nanometer (40 nm) CMOS. It demonstrates that the performance of TD sensors indeed continues to improve with scaling. Without temperature calibration, the sensor achieves $\pm 1.4\text{ }^{\circ}\text{C}$ (3σ) inaccuracy from $-40\text{ }^{\circ}\text{C}$ to $125\text{ }^{\circ}\text{C}$, which is $5\times$ better than previous (non-TD) sensors intended for thermal monitoring [4], [14]–[16]. This improves to $\pm 0.75\text{ }^{\circ}\text{C}$ (3σ) after a single-temperature calibration, a level of accuracy that, for non-TD sensors, requires two-temperature calibration [4], [14], [15]. Furthermore, it operates from a 0.9-V supply, and occupies only $1650\text{ }\mu\text{m}^2$, making it one of the smallest smart temperature sensors reported to date.

This paper begins with a description of the ETF design in Section II and continues with the system level design in Section III. The circuit implementation is detailed in Section IV. Experimental results are shown in Section V and the conclusions are drawn in Section VI.

II. ELECTROTHERMAL FILTER DESIGN

The simplified layout of an ETF realized in a standard 40-nm CMOS process is shown in Fig. 1. The heater is a diffusion resistor, while the relative temperature sensor is a

thermopile, i.e., a series connection of p+ silicon/aluminum thermocouples. The heater is driven by a square wave at a constant frequency, so that the ETF's temperature-dependent delay manifests itself as a phase shift in the thermopile's output voltage. The whole structure is placed in an n-well to shield it from electrical interference via the substrate. The effect of thermal interference via the substrate (e.g., due to other on chip circuitry) is not a concern, since this will be strongly low-pass filtered in the thermal domain [16].

In Fig. 1, the hot junction of each thermocouple, i.e., the p+/Al contact closest to the heater, is located at a distance s from the ETF's center, while the cold junctions are further away. Since each thermocouple produces a voltage proportional to the temperature difference between its hot and cold junctions, the ETF's output signal is larger for a shorter s and for longer thermocouple arms. However, reducing s means a larger sensitivity to lithographic errors, thus resulting in lower accuracy, while longer thermocouple arms have higher resistance, thus causing higher thermal noise.

Previous ETFs were optimized for accuracy at the expense of signal-to-noise ratio (SNR), which meant that their heater/thermopile spacing was relatively large ($s = 24\text{ }\mu\text{m}$). As a result, their readout bandwidth had to be less than 1 Hz to achieve reasonable resolution [17]. In this paper, we leverage the improved lithographic accuracy of nanometer CMOS to implement ETFs with much smaller heater/thermopile spacing in order to improve SNR without significantly degrading accuracy. Moreover, an octagonal layout is used that minimizes thermopile resistance, and hence thermal noise, by maximizing the thermopile width. In this paper, two ETFs were realized, with $s = 3.3$ and $2\text{ }\mu\text{m}$, respectively, in order to explore the influence of s on ETF performance. Both ETFs occupy an area of $240\text{ }\mu\text{m}^2$ and dissipate an average power of 2.1 mW from a 1.05-V supply.

For compatibility with previous work [12], the ETF drive frequency (F_{DRIVE}) is set at 1.17 MHz . From $-40\text{ }^{\circ}\text{C}$ to $125\text{ }^{\circ}\text{C}$, the phase shifts of the $s = 3.3\text{-}$ and $2\text{-}\mu\text{m}$ ETFs are then expected to range from 35° to 60° , and from 25° to 45° , respectively. Based on thermal modeling, the corresponding output levels are expected to be 1.3 and $2.4\text{ mV}_{\text{pp}}$, respectively, for a heater power dissipation of 1 mW [18]. Combined with the parasitic capacitance of the thermopiles, the thermopile's resistance R_{TP} , about 8 and $12\text{ k}\Omega$ for the 3.3- and $2\text{-}\mu\text{m}$ ETFs, respectively, causes an additional phase shift of 0.4° and 0.6° . The spread on this RC phase shift (about 30% over corners) will give rise to an equivalent temperature-sensing spread of less than $0.9\text{ }^{\circ}\text{C}$.

III. SYSTEM LEVEL DESIGN

The target sampling rate of 1 kSa/s and the small area requirement pose significant challenges on the design of the readout architecture. Fundamentally, an ETF's temperature information is contained in the phase delay of a small ($\sim\text{mV}$ amplitude) signal, and so a sensitive and high-resolution phase-domain ADC is necessary. As shown in previous work [11], [17], the phase-domain $\Sigma\Delta$ modulator (PD $\Sigma\Delta$ M) is a good candidate for this purpose. A PD $\Sigma\Delta$ M is the $\Sigma\Delta$ modulator with a feedback path in the

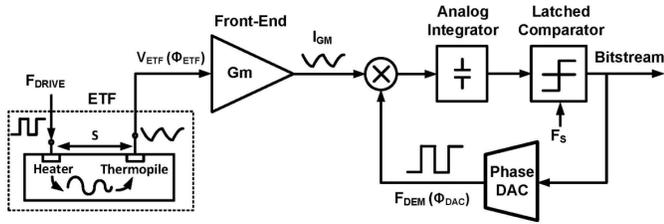


Fig. 2. Simple block diagram of a phase-domain $\Sigma\Delta$ digitizing the phase output of an ETF.

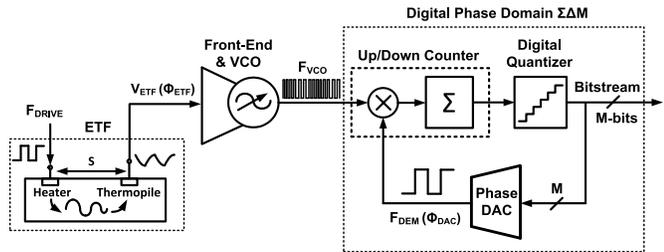


Fig. 3. Block diagram of a VCO-based phase-domain $\Sigma\Delta$.

phase domain. The required phase-domain summation node can be realized by a chopper demodulator, which demodulates the phase of the ETF signal (at a frequency F_{DRIVE}) by multiplying it with a square wave at F_{DRIVE} , but with a known (reference) phase shift [17]. A PD $\Sigma\Delta$ thus incorporates a synchronous phase detector and as such is only sensitive to interferers at frequencies very close to the drive frequency F_{DRIVE} . In an SoC, the presence of such interferers can readily be avoided by proper frequency planning.

Fig. 2 shows the block diagram of a first-order PD $\Sigma\Delta$. A gm-stage converts the ETF's output voltage (at frequency F_{DRIVE}) into current, whose phase shift (Φ_{ETF} , measured with respect to the phase of the signal driving the ETF's heater) is detected by a chopper demodulator driven by F_{DEM} . The phase-dependent dc current is then integrated on a capacitor and applied to a latched comparator, whose bitstream (BS) output switches F_{DEM} between outputs of a phase DAC (Φ_{DAC}) in the $\Sigma\Delta$ manner. For a single-bit modulator, Φ_{DAC} switches between the two phase references, Φ_0 and Φ_1 . F_{DRIVE} and phase DAC outputs can be generated by a digital block, which is driven by an accurate high-frequency clock (F_{SYNC}) [12]. However, such PD $\Sigma\Delta$ s require large integration capacitors and high-gain amplifiers [12], which, in turn, occupy significant area. Moreover, because of the need for high-gain amplifiers, this architecture does not scale well with technology [19].

A more digital-friendly architecture was proposed in [20] and is shown in Fig. 3. In such voltage-controlled oscillator (VCO)-based PD $\Sigma\Delta$, the combination of a VCO and an up/down counter replaces the gm-stage, the chopper, and the integration capacitor. Here, the ETF output signal V_{ETF} at frequency F_{DRIVE} and phase shift Φ_{ETF} modulates the VCO's output frequency (F_{VCO}). An all-digital $\Delta\Sigma$ modulator then synchronously demodulates the VCO's output and digitizes the ETF's phase shift Φ_{ETF} . The functions of demodulation and integration are realized by the up/down counter, whose M most significant bits (MSBs) of its output word constitute

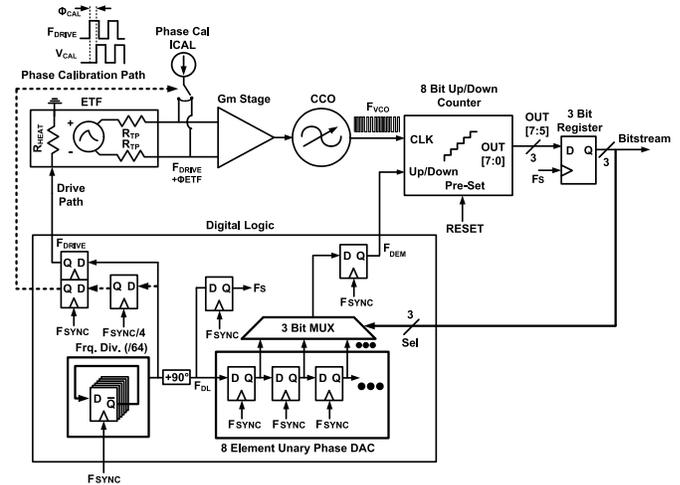


Fig. 4. Block diagram of the 3-b VCO-based PD $\Sigma\Delta$ M with phase calibration.

the output BS. Such BS drives the phase DAC, which applies a digitally delayed feedback signal (F_{DEM}) to the counter's up/down input. To improve accuracy, the modulator is usually operated as an incremental converter, where the counter is reset before each conversion [21]. The decimation filter can then be a simple counter (sinc filter) [20]. In contrast to previous work, a multibit DAC ($M = 3$) is chosen in this paper, a choice which reduces both quantization noise and the inherent cosine nonlinearity of synchronous phase demodulation [11] to negligible levels ($\pm 0.04^\circ\text{C}$). This avoids the complexity of a two-step conversion with single-bit incremental converters [12], without compromising performance.

However, a disadvantage of the proposed architecture is that the finite bandwidth of the VCO in Fig. 3 results in additional phase shift, which cannot be distinguished from Φ_{ETF} . In fact, while the gm-stage in Fig. 2 can be implemented by a fast differential pair immediately followed by a demodulating analog chopper [17], the VCO in Fig. 3 requires both a low-noise front-end and a cascaded oscillator element, and thus is inherently slower. In this paper, the VCO's phase error is mitigated by a phase-calibration scheme in which the entire VCO-based PD $\Sigma\Delta$ M is driven by a reference square wave (V_{CAL}) with a known phase shift (Φ_{CAL}). The additional phase error introduced by the readout can thus be determined and then subtracted from the results of subsequent conversions.

Fig. 4 shows the block diagram of a VCO-based PD $\Sigma\Delta$ M with phase calibration. Here, the VCO front-end is implemented as a gm-stage followed by a current-controlled oscillator (CCO). The gm-stage isolates the weak $\sim mV_{pp}$ ETF signal from the CCO to prevent kickback and also acts as a low-noise amplifier. The CCO drives an 8-b up-down counter, whose 3 MSBs are latched by D flip-flops to realize the quantizer of a 3-b $\Sigma\Delta$ modulator. The 3-b unary phase DAC consists of a 3-b multiplexer selecting the outputs of an eight-element delay line that shifts an input signal (F_{DL}), where $\angle F_{DL} = \angle F_{DRIVE} + 90^\circ$. The reference delay signal (F_{SYNC}) is an external 75-MHz clock, while $F_{DRIVE} = 1.17$ MHz. This results in a phase DAC LSB of 5.625° , but in order to cover a large range, the DAC LSB was chosen to be 11.25° , in practice, via dividing F_{SYNC} by 2. Therefore, the

TABLE I
NOISE, DELAY, AND POWER BUDGETING BETWEEN ETF
AND READOUT BLOCKS

Circuit Block	Thermal Noise Density (Voltage)	Noise Density* (Phase)	Power**	Phase Delay ($F_{DRIVE} = 1.17$ MHz)
ETF ($s = 2 \mu\text{m}$)	13.7 nV/ $\sqrt{\text{Hz}}$	1.01 $\text{m}^2/\sqrt{\text{Hz}}$	2.1 mW	0.6°
ETF ($s = 3.3 \mu\text{m}$)	11.4 nV/ $\sqrt{\text{Hz}}$	1.54 $\text{m}^2/\sqrt{\text{Hz}}$	2.1 mW	0.4°
Gm-Stage + CCO ($s = 2 \mu\text{m}$)	10 nV/ $\sqrt{\text{Hz}}$	0.73 $\text{m}^2/\sqrt{\text{Hz}}$	0.17 mW	0.75°
Gm-Stage + CCO ($s = 3.3 \mu\text{m}$)		1.35 $\text{m}^2/\sqrt{\text{Hz}}$		
Up/Down Counter	-	-	0.26 mW	-
Phase DAC	-	-	0.01 mW	0.1°
Total ($s = 2 \mu\text{m}$)	17 nV/ $\sqrt{\text{Hz}}$	1.24 $\text{m}^2/\sqrt{\text{Hz}}$	2.5 mW	1.45°
Total ($s = 3.3 \mu\text{m}$)	15.2 nV/ $\sqrt{\text{Hz}}$	2.05 $\text{m}^2/\sqrt{\text{Hz}}$	2.5 mW	1.25°

* 1.3-mVpp ETF signal assumed for voltage to phase noise conversion for $s = 3.3 \mu\text{m}$

** 2.4-mVpp ETF signal assumed for voltage to phase noise conversion for $s = 2 \mu\text{m}$

** $V_{DD} = 1.05$ V

DAC spans from 101.25° to 180° . In order to minimize any circuit related delay and, hence, any additional phase error in F_{DRIVE} and F_{DEM} , both clock signals are synchronized by F_{SYNC} before being delivered to the heater switches or to the up/down counter. Unlike prior work employing analog choppers [12], low-frequency chopping is not necessary to eliminate the residual offset due to chopper nonidealities, because the up/down counter behaves like a near-ideal digital chopper. This further simplifies the drive logic, thus saving additional area.

The phase-calibration reference signal is generated by injecting a reference current from a current source (I_{CAL}) into the thermopile's resistances R_{TP} . The reference phase for phase calibration was chosen equal to 22.5° , a phase which requires only two flip-flops to generate.

The total budget for thermal noise (resolution), electrical phase delay (accuracy), and power of the proposed TD sensor is shown in Table I. The gm-stage is optimized for low-power consumption and low area, thus leading to a gm-stage design that contributes $\sim 30\%$ of the total thermal noise and about half of the phase delay budgets.

In addition to thermal noise, the $\text{PD}\Sigma\Delta\text{M}$'s resolution is also affected by the quantization noise imposed by the CCO and counter combination. This occurs, because the counter only recognizes the rising edges of F_{VCO} , effectively quantizing the time-domain information coming from the CCO. In most amplitude-domain VCO-based ADCs, the VCO is cascaded to a fully analog loop filter, thus providing high-pass shaping of this noise and effectively removing it from the band of interest [22]. Unfortunately, this is not the case for a phase-domain modulator. Indeed, CppSim [23] simulations of the $\text{PD}\Sigma\Delta\text{M}$ shown in Fig. 4 reveal that this quantization noise manifests itself as an input-referred white noise source. Nevertheless, simulations also confirm intuition in showing that such time-domain quantization noise is lower for a higher F_{VCO} frequency. For the proposed design, the nominal VCO frequency (F_{NOM}) is 630 MHz, while the voltage-to-frequency gain of the VCO (K_{VCO}) is 200 MHz/mV. With these values, simulations show that the additional quantization noise due to

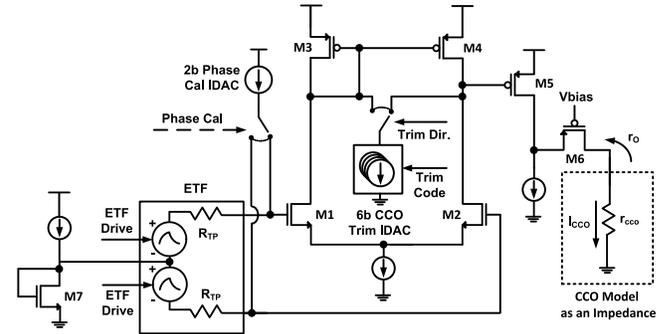


Fig. 5. Circuit diagram of the gm-stage (cascaded CCO modeled as resistive load r_{CCO}).

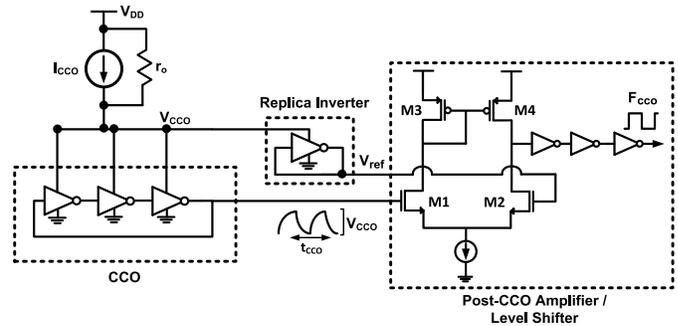


Fig. 6. Circuit diagram of the CCO and the cascaded level shifter amplifier. The driving gm-stage is modeled with its Norton equivalent (I_{CCO} current source and r_o output impedance).

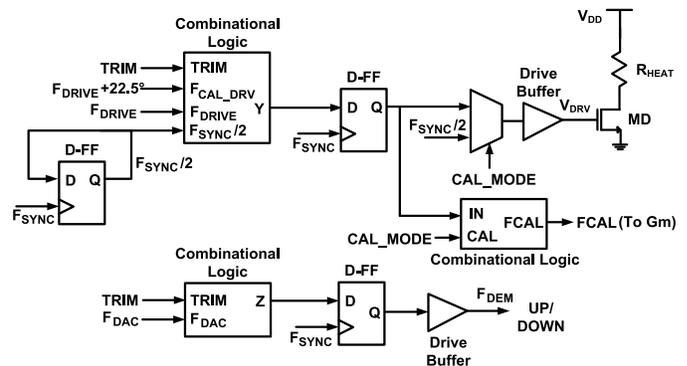


Fig. 7. Block diagram of the sensor digital logic for generation of F_{DRIVE} and F_{DEM} and truth table that describes the combinational logic function.

VCO is about 25 m° in a 500-Hz bandwidth, which translates into a temperature-sensing resolution of $0.16 \text{ }^\circ\text{C}$.

IV. CIRCUIT DESCRIPTION

Fig. 5 shows the circuit level implementation of the gm-stage that supplies the CCO. The CCO can be modeled as a nonlinear impedance r_{CCO} sinking a current I_{CCO} . For maximum efficiency and driving capability, $r_{CCO} \ll r_o$, where r_o is the output impedance of the gm-stage. Although r_{CCO} depends on the CCO architecture, it is typically in the order

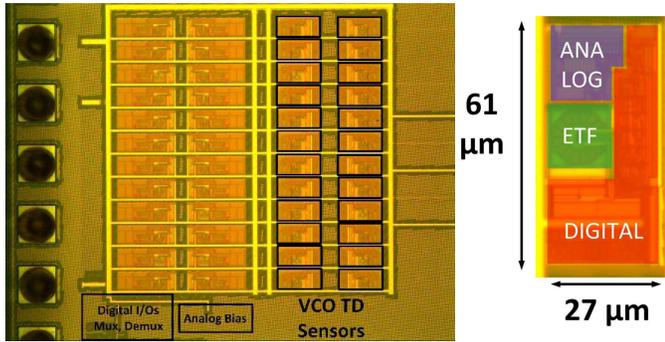


Fig. 8. Die photo, along with a zoomed-in photo of a single temperature sensor. The sensor's photo is showing the breakdown of area occupied by the ETF and circuitry.

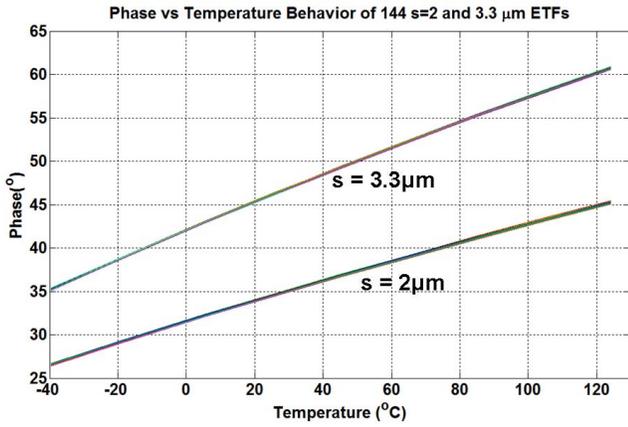


Fig. 9. Measured phase of $s = 2\text{-}\mu\text{m}$ and $s = 3.3\text{-}\mu\text{m}$ ETFs over temperature ($F_{\text{DRIVE}} = 1.17\text{ MHz}$ and 144 samples).

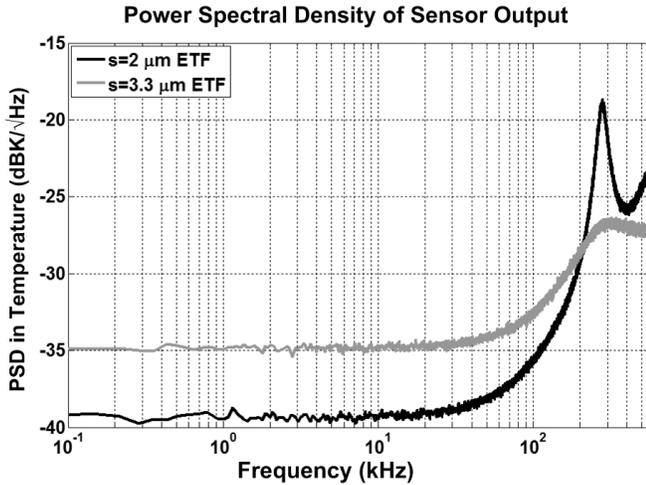


Fig. 10. PSD of the sensor's BS (eight million points and $F_s = 1.17\text{ MHz}$).

of tens of kilohms. Therefore, the gm-stage requires a high output impedance, as well as a high transconductance (gm) to meet the noise requirements shown in Table I. Moreover, it needs to work a supply voltage below 1 V to demonstrate compliance with current and future supply voltages for nanometer CMOS (1.1 V for 40-nm CMOS). In 40-nm CMOS technology, these three requirements necessitate the use of a two-stage amplifier architecture. A two-stage design also uses

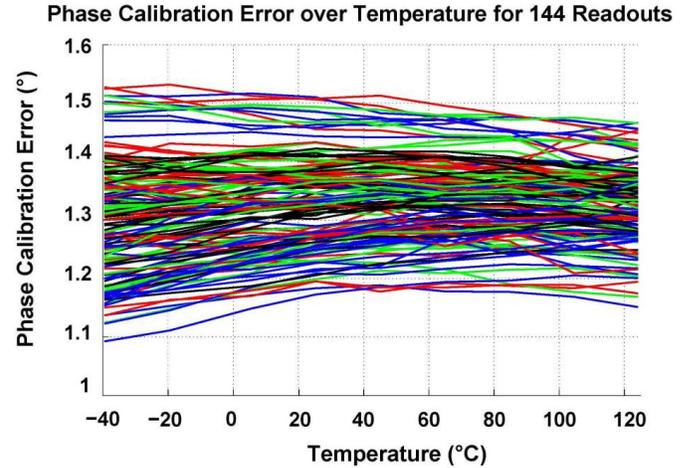


Fig. 11. Measured phase error of the readout circuitry of 144 sensor readouts from $-40\text{ }^{\circ}\text{C}$ to $125\text{ }^{\circ}\text{C}$.

less transistors (8) than a folded-cascode (11) amplifier [17], and thus occupies less area. Although a two-stage amplifier may have larger delay than a single-stage amplifier, this can be compensated by the phase calibration.

The first stage (M1-4) is optimized for minimal thermal noise, and phase shift at F_{DRIVE} and has a gain of 25 dB and a bandwidth of 300 MHz. Its $10\text{-nV}/\sqrt{\text{Hz}}$ noise density (see Table I) is mostly dominated by the input pair M1-2. The second stage (M5) adds gain for an overall gm of 2.5 mA/V. It is cascoded by M6 to boost its output impedance r_O from ~ 80 to $\sim 400\text{ k}\Omega$ without significantly compromising CCO's voltage headroom. With this configuration, the circuit operates correctly with a supply voltage as low as $2 V_{\text{GS}} + 2 V_{\text{DS}} \cong 0.8\text{ V}$ ($2 V_{\text{GS}}$ for the CCO headroom and $2 V_{\text{DS}}$ for M5 and M6 in Fig. 5).

The offset of the gm-stage together with the PVT variations of the CCO can create a large spread in the nominal CCO frequency F_{NOM} . An excessively high F_{NOM} can cause counter failure while an excessively low F_{NOM} can both cause excessive quantization noise and force the CCO in a highly nonlinear operating region. Moreover, large changes of F_{NOM} over temperature can cause the delay of the VCO to change, and add a temperature-dependent phase error, i.e., more inaccuracy. Therefore, F_{NOM} is trimmed by a 6-b current DAC (IDAC) before every conversion. During this process, the counter is configured to only count up, while external logic implements a simple ramp algorithm that monitors the counter's fourth LSB (toggling at $F_{\text{VCO}}/16$) and increments the IDAC's input until F_{VCO} is $\sim 630\text{ MHz}$. This whole calibration process takes less than $100\text{ }\mu\text{s}$ over the specified supply voltage and temperature range. One LSB of the trimming IDAC corresponds to a 62.5-MHz average step on F_{NOM} , thus resulting in $F_{\text{NOM}} = 630\text{ MHz} \pm 62.5\text{ MHz}$, which is enough to guarantee negligible phase error. The IDAC can compensate an error up to $\pm 20\text{ mV}$ referred at the gm-stage input, which is large enough to cover PVT variations as well as amplifier offset.

During phase calibration, the current source for phase calibration (I_{CAL} in Fig. 4) is switched between the two thermopile resistors of the ETF to generate an ac square wave

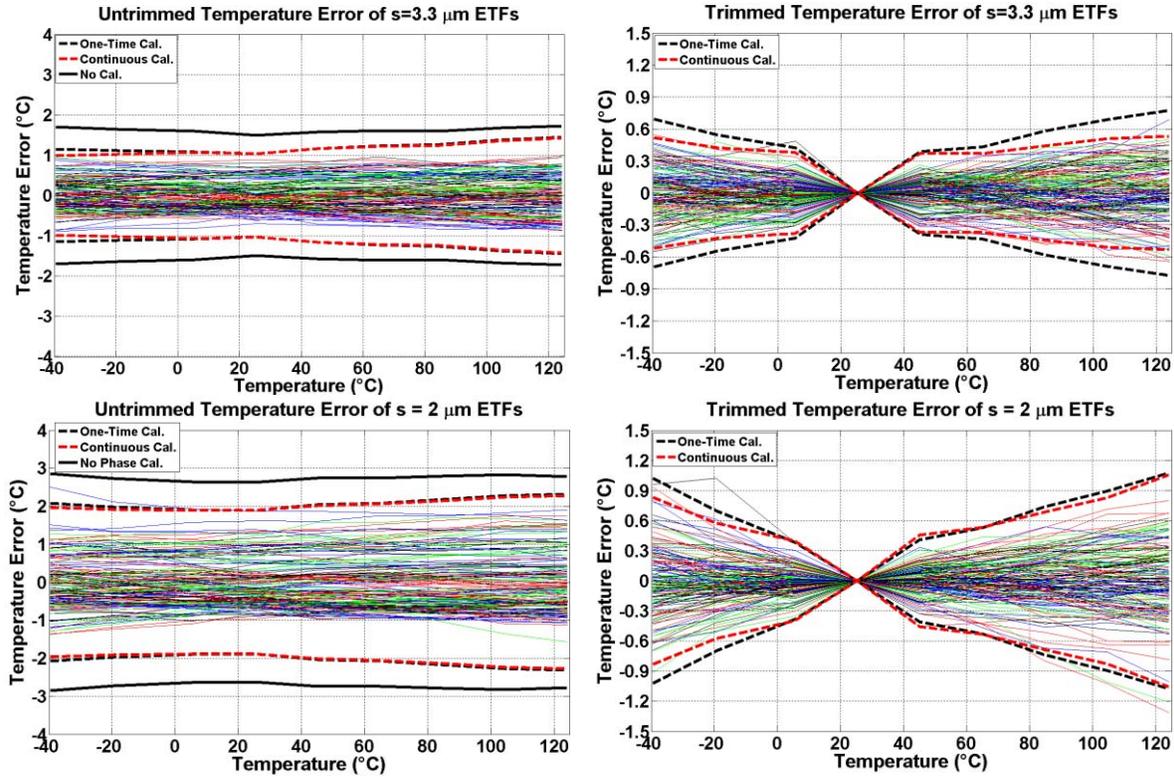


Fig. 12. Untrimmed and single-point trim inaccuracy for 144 sensors with $s = 3.3 \mu\text{m}$ (top plots) and $s = 2 \mu\text{m}$ (bottom plots). Individual lines represent the inaccuracy of each sensor with one-time phase cal., while the bold lines indicate the 3σ limits for no phase cal., one-time phase cal. at 25°C , and the red dashed lines represent continuous phase cal.

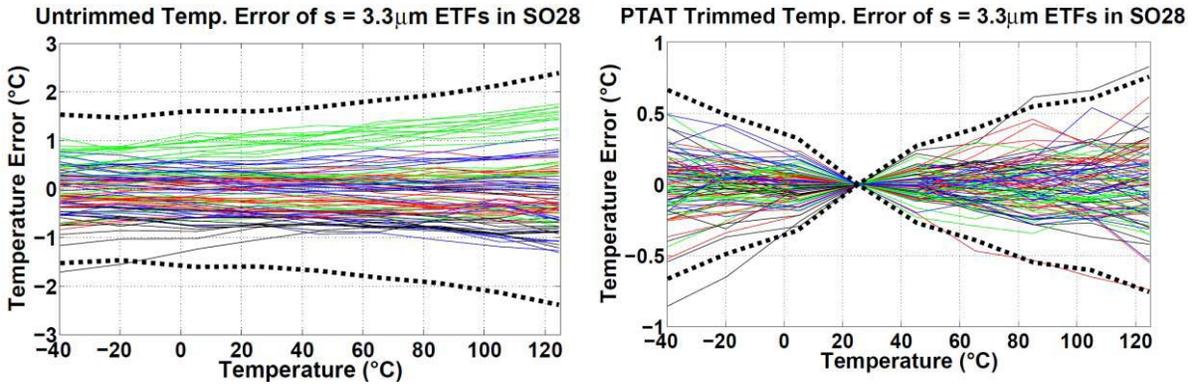


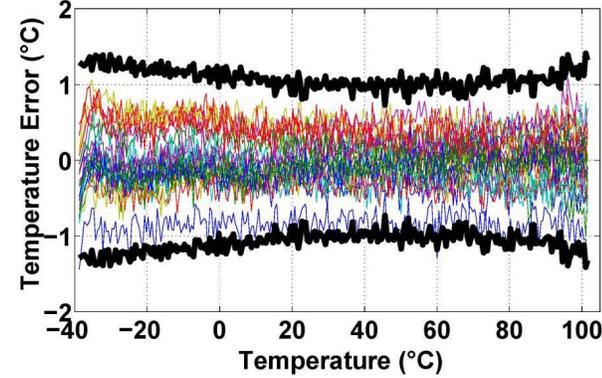
Fig. 13. Untrimmed and gain or PTAT trimmed inaccuracy for 96 sensors with $s = 3.3 \mu\text{m}$, in 16 SO28 plastic packages. Individual lines represent the inaccuracy of each sensor with one-time phase cal., while the dashed lines indicate the 3σ limits for one-time phase cal. at 25°C .

with an amplitude up to 2 mV_{pp} at the gm-input. Biasing transistor M7 determines the common mode voltage of the ETF thermopiles. I_{CAL} has been designed as a 2-b current DAC with a unit current of 125 nA , in order to test the effect of front-end nonlinearity on the phase-calibration technique. Since this nonlinearity was found to be negligible during experimental characterization, I_{CAL} is always operated at its maximum current of 500 nA .

Fig. 6 shows the circuit level implementation of CCO. The gm-stage is modeled as a current source I_{CCO} with impedance source resistance of r_O . In order to minimize area and maximize CCO gain (K_{CCO}), the CCO is implemented as a ring oscillator with the minimum number of stages,

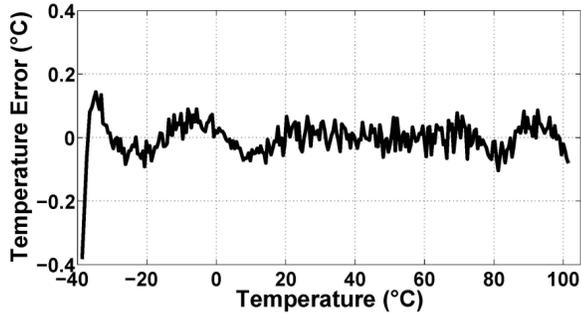
i.e., three stages. Each transistor in the inverters is sized with minimum length and twice the minimum width, to ensure that CCO output swing is low enough to ensure voltage headroom at the output of the gm-state for sub-1-V operation. With these design choices, r_{CCO} is $\sim 30 \text{ k}\Omega$ at 25°C and is much smaller than r_O , as intended.

The impact of the CCO's phase noise on the sensor's resolution is reduced by the gain of the preceding gm-stage. Moreover, only a narrowband component of this noise around F_{DRIVE} is involved, since the CCO's output is synchronously demodulated. As a result, the noise of the gm-stage is dominant in this design. As explained before, the CCO's PVT variation is corrected by trimming F_{NOM} before every conversion.

Temp. Error of 24 s = 3.3 μm ETFs During Ramp Meas.

(a)

Mean Non-Linearity Error Compared to Oven Ramp



(b)

Fig. 14. (a) Temperature error of 24 sensors with 3.3- μm ETFs during a ramped temperature test (50-mK/sample temperature slope and 1-kSa/s sample rate). Bold lines indicate 3σ limits. (b) Nonlinearity error between oven ramp and mean sensor output over temperature.

Since the CCO's voltage swing is small and depends on PVT, it is boosted to logic-compatible levels by a single-stage differential amplifier (M1-4) referenced to a replica inverter. The amplifier is designed for speed, since its delay adds to the phase shift of the gm-stage. It has a nominal ac gain of 10 dB over a bandwidth of 900 MHz and consumes only 50 μA . After the amplifier, three tapered inverters provide the strength to drive the 8-b counter.

The 8-b up/down counter was synthesized from standard cells and laid out via a standard place-&-route tool. According to simulation, it can operate at a clock frequency up to 1 GHz over all corners, 0.9–1.2-V supply voltage, and the temperature range. The up/down signal (F_{DEM}) is relocked by F_{VCO} to avoid metastability in the counter. The 3-b sampling register also employs standard cells and its sampling clock (F_{S}) is generated by the digital logic and is relocked by the falling edge of F_{VCO} , which means that the up/down counter must settle within a half period of F_{VCO} .

Fig. 7 shows the schematic of the digital logic that generates the signals driving the ETF heater and the counter, along with the truth table describing the function of the combinational logic. The heater-driving transistor MD controls the current flow in the ETF heater R_{HEAT} to create the ETF heat pulse. To minimize the parasitic series resistance and, hence, maximize the power efficiency of the ETF, each heater ($R_{\text{HEAT}} = 188 \Omega$) is driven by a large nMOS

($W = 68 \mu\text{m}$, $L = 40 \text{ nm}$, and $R_{\text{on}} \sim 20 \Omega$). The large gate capacitance of MD is driven by a digital buffer implemented as tapered inverters. Since any delay mismatch between F_{DRIVE} and F_{DEM} would result in a phase error and, consequently, in additional inaccuracy, the up/down signal path mirrors the drive path by using the same synchronizing flip-flop and digital buffer between the phase DAC output (F_{DAC}) and the counter input (F_{DEM}).

The signals CAL_MODE and TRIM set the system in phase calibration and CCO trimming modes, respectively. When either mode is selected, a relatively high frequency signal ($F_{\text{SYNC}}/2$) is provided to the ETF. At this frequency, the ETF's ac output is quite small, while the same self-heating-induced dc offset is present as in normal operation [11]. In addition, when phase calibration is enabled, a delayed version of F_{DRIVE} (generated by an auxiliary output of the phase DAC) is delivered to the gm-stage via F_{CAL} . When TRIM mode is enabled, the counter is forced to count only up and both the ETF input, and the F_{CAL} signals are disabled to guarantee that the VCO only sees the offset of the gm-stage and the self-heating of the ETF.

V. EXPERIMENTAL RESULTS

The prototype was realized in a standard 40-nm CMOS process and occupies an active area of 0.23 mm² (Fig. 8). It consists of an array of 12 \times 2 sensors, 12 \times 2 test structures, two test heaters (resistors), a shared bias-current generator, and shared digital I/O logic (shift registers and multiplexers for testability). Each sensor occupies 61 $\mu\text{m} \times 27 \mu\text{m}$, and dissipates 2.5 mW, most of which (88%) is dissipated in the ETF.

In each sensor, the ETF occupies only 15% of the 1650- μm^2 sensor area, while the analog and digital circuitries occupy 25% and 60%, respectively. In 40-nm CMOS, the sensor is about 2 \times smaller than previous designs in 160-nm CMOS [13], even though it includes many additional features, such as phase calibration, multibit feedback, and the phase DAC's reference generation. The area required for the decimation filter and the CCO's trimming logic is estimated to be about 600 μm^2 , but since those functions do not necessarily need to be colocated within the sensor, they were implemented off-chip for flexibility. Functionality of each sensor was verified over (digital and analog) supply voltages ranging from 0.9 to 1.2 V (nominal supply is 1.05 V), and a 2.8 $^{\circ}\text{C}/\text{V}$ supply sensitivity was observed over such range.

The phase versus temperature characteristics of both ETFs from -40°C to 125 $^{\circ}\text{C}$ (averaged over 24 dies and 144 sensors for each ETF) at 1.17-MHz drive frequency were used to generate the fifth-order polynomial master curves shown in Fig. 9. Those master curves were used to convert the decimated output of each PD Σ ΔM into a temperature reading. Over the measured temperature range, the master curves can be well approximated by a T^n power law [23]. For the 3.3- and 2- μm ETFs, good fits were obtained with $n = 0.98$ and $n = 0.95$, respectively, which agrees well with previous work [12], [24].

Fig. 10 shows the power spectral density (PSD) of the 3-b digital output of both the 3.3- and 2- μm ETFs. The thermal noise floor corresponds to a resolution of 0.36 $^{\circ}\text{C}$ (rms) for

TABLE II
PERFORMANCE SUMMARY AND COMPARISON

	This Work		[13]	[15]	[5]	[4]	[26]	[27]	[14]
Technology	40nm		160nm	32nm	14nm	32nm	16nm	28nm	65nm
Sensor Type	TD (3.3 μ m)	TD (2 μ m)	TD (3.3 μ m)	Diode	BJT	BJT	BJT	BJT	MOS
Inaccuracy No Temp Cal. (3 σ , °C)	± 1.4	± 2.3	± 2.9	-	± 4.7	± 5	± 2.0	± 1.8	-
Single Temp. Cal. (3 σ , °C)	± 0.75	± 1.05	± 1.2	-	± 2.3	-	-	-	-
Two Temp. Cal. (3 σ , °C)	-	-	-	± 2.6	± 0.7	-	-	-	$\pm 0.9^*$
Temp. Range (°C)	-40 to 125	-40 to 125	-35 to 125	0 to 100	0 to 100	-10 to 110	-50 to 150	-20 to 130	0 to 100
Area (μ m ²)	1650		2800***	1000**	8700	20000	12600	3800	4000
Resolution (°C, RMS)	0.36	0.24	0.47	0.25	0.5	0.15	0.38	0.58	0.3
Speed (kSa/s)	1		1	2.5	50	1.2	3.66	250	45
Supply Voltage (V)	0.9 – 1.2		1.8	1.65	1.35	1.05	-	1.1 - 1.8	0.85-1.05
Power (mW)	2.5		2.4	0.1	1.1	1.6	1.21	0.016	0.15
Resolution FoM (nJ·K ²)****	324	144	530	2.5	5.5	30	47	0.021	0.3

* Peak to peak error variation (7 samples)

** Area of precision voltage reference not included

*** Shared phase DAC area (~600 μ m²) not included

**** Resolution figure of merit (FoM) is defined as Power*Conversion Time*Resolution²

the 3.3- μ m ETF and 0.24 °C (rms) for the 2- μ m ETF, both obtained for a bandwidth of 500 Hz, i.e., at a sampling rate of 1 kSa/s.

The additional phase due to the readout can be detected and removed via phase calibration. Fig. 11 shows the phase error of the readout circuitry of 144 sensors, measured at a reference phase of 22.5° (Fig. 11). The mean phase error is 1.3° and it exhibits a slight curvature over temperature. Phase calibration can be done continuously, e.g., after every conversion, but at the expense of halving the conversion rate and degrading the resolution from 0.24 °C for the 2- μ m ETF (0.36 °C for the 3.3- μ m ETF) to 0.40 °C (0.5 °C). Alternatively, it can be done one time at room temperature after fabrication but at the expense of increased inaccuracy.

As shown in Fig. 12, the sensors based on the 3.3- μ m ETF achieve an untrimmed inaccuracy of ± 1.8 °C (3 σ , 144 sensors, and 24 dies) from -40 °C to 125 °C for a supply voltage of 1.05 V. The inaccuracy improves to ± 1.4 °C (3 σ) after a one-time phase calibration at room temperature, and to ± 0.75 °C (3 σ) after temperature calibration at 25 °C. Continuous phase calibration improves inaccuracy to ± 0.5 °C (3 σ). At a 0.9 V supply voltage, the digital logic slows down, resulting in an untrimmed inaccuracy of ± 2.3 °C (3 σ), and ± 1.2 °C (3 σ) after trimming.

The improved resolution of the 2- μ m ETFs comes at the expense of accuracy, as shown in Fig. 12. Their untrimmed inaccuracy is ± 2.3 °C (3 σ , 144 sensors, and 24 dies) after a one-time or continuous phase calibration. After a

single-temperature calibration, those values reduce to ± 1.05 °C (3 σ) and ± 0.85 °C (3 σ), respectively.

Self-heating of the ETFs (1.7 °C and 4 °C for the 3.3- and 2- μ m ETF, respectively) is estimated to spread by approximately 20% due to the spread in heater resistance and in parasitic resistance of the driving transistor. This results in an error of ± 0.35 °C (3 σ) for the 3.3- μ m ETF and ± 0.8 °C (3 σ) for the 2- μ m ETF, which is already included in the ± 1.4 °C (3.3- μ m ETF) and ± 2.3 °C (2- μ m ETF) values reported earlier.

In order to test the sensor's sensitivity to mechanical stress, 16 dies (each containing 6 \times 3.3- μ m ETF and 6 \times 2- μ m ETFs) were packaged in standard SO28 plastic packages. As shown in Fig. 13, the untrimmed inaccuracy of 96, 3.3- μ m ETFs, was ± 2.3 °C (3 σ). Compared with the ceramic-packaged devices, more spread was observed, which may be due to the additional self-heating in plastic packages and to the stress sensitivity of the TD of silicon [25]. After a PTAT trim, the spread drops to ± 0.75 °C (3 σ), which is the same for plastic and ceramic packaged sensors.

To characterize the nonlinearity of the PD Σ Δ M, they were exposed to a temperature ramp from -40 °C to 125 °C. Fig. 14 shows the statistical averages obtained from a 50-mK/sample ramp. It can be seen that no artifacts occur during the measurement.

The sensor's performance with both ETFs is summarized in Table II and compared with that of other sensors intended for thermal-monitoring applications. Due to the amount of

power dissipated in the ETF, the proposed sensor is not particularly energy efficient, as can be seen from its relatively poor resolution FoM [28]. However, with the 3.3- μm ETF, the proposed sensor is the most accurate and the smallest, except for a sensor that requires an accurate external voltage reference (which is not included in the reported area) [15]. It also has the second lowest operating supply voltage (0.9 V), which is mainly limited by the up/down counter. Compared with TD sensors implemented in more mature technologies [13], it achieves $1.5\times$ better resolution and $2\times$ more accuracy, while requiring about $2\times$ less area.

VI. CONCLUSION

A compact TD sensor in 40-nm CMOS has been described, and techniques, which allow the sensor to be implemented in a compact area, have been presented. The sensor's area, speed, resolution, and power-supply rejection satisfy typical specifications for SoC thermal monitoring, while its untrimmed inaccuracy is the lowest reported for temperature sensors in nanometer CMOS below 40 nm. The performance (area, accuracy, power, and speed) of TD sensors has been demonstrated to improve with process scaling, and additional improvements can be reasonably expected in the future more advanced technologies. These results demonstrate that the TD-based temperature sensors are suitable for hotspot monitoring in microprocessors and other systems-on-chip.

REFERENCES

- [1] E. Rotem, J. Hermerding, C. Aviad, and C. Harel, "Temperature measurement in the Intel Core Duo processor," in *Proc. THERMINIC*, Sep. 2006, pp. 23–27.
- [2] J. S. Shor and K. Luria, "Miniaturized BJT-based thermal sensor for microprocessors in 32- and 22-nm technologies," *IEEE J. Solid-State Circuits*, vol. 48, no. 11, pp. 2860–2867, Nov. 2013.
- [3] M. Floyd *et al.*, "Introducing the adaptive energy management features of the POWER7 chip," *IEEE Micro*, vol. 31, no. 2, pp. 60–75, Mar. 2011.
- [4] H. Lakdawala, Y. W. Li, A. Raychowdhury, G. Taylor, and K. Soumyanath, "A 1.05V 1.6mW 0.45 °C 3 σ -resolution $\Delta\Sigma$ -based temperature sensor with parasitic-resistance compensation in 32nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 12, pp. 3621–3630, Dec. 2009.
- [5] T. Oshita, J. Shor, D. E. Duarte, A. Kornfeld, and D. Zilberman, "Compact BJT-based thermal sensor for processor applications in a 14 nm tri-gate CMOS process," *IEEE J. Solid-State Circuits*, vol. 50, no. 3, pp. 799–807, Mar. 2015.
- [6] F. Sebastiano, L. J. Breems, K. A. A. Makinwa, S. Drago, D. M. W. Leenaerts, and B. Nauta, "A 1.2-V 10- μW NPN-based temperature sensor in 65-nm CMOS with an inaccuracy of 0.2 °C (3σ) from -70 °C to 125 °C," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2591–2601, Dec. 2010.
- [7] J.-J. Horng *et al.*, "A 0.7V resistive sensor with temperature/voltage detection function in 16nm FinFET technologies," in *VLSI Symp. Dig.*, Jun. 2014, pp. 1–2.
- [8] D. Ha, K. Woo, S. Meninger, T. Xanthopoulos, E. Crain, and D. Ham, "Time-domain CMOS temperature sensors with dual delay-locked loops for microprocessor thermal monitoring," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 9, pp. 1590–1601, Sep. 2012.
- [9] S. Hwang, J. Koo, K. Kim, H. Lee, and C. Kim, "A 0.008 mm² 500 μW 469 kS/s frequency-to-digital converter based CMOS temperature sensor with process variation compensation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 9, pp. 2241–2248, Sep. 2013.
- [10] K. A. A. Makinwa and M. F. Snoeij, "A CMOS temperature-to-frequency converter with an inaccuracy of less than ± 0.5 °C (3σ) from -40 °C to 105 °C," *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2992–2997, Dec. 2006.
- [11] C. P. L. van Vroonhoven, D. d'Aquino, and K. A. A. Makinwa, "A thermal-diffusivity-based temperature sensor with an untrimmed inaccuracy of ± 0.2 °C (3σ) from -55 °C to 125 °C," in *ISSCC Dig. Tech. Papers*, Feb. 2010, pp. 314–315.
- [12] U. Sönmez, R. Quan, F. Sebastiano, and K. A. A. Makinwa, "A 0.008-mm² area-optimized thermal-diffusivity-based temperature sensor in 160-nm CMOS for SoC thermal monitoring," in *Proc. 40th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2014, pp. 395–398.
- [13] J. Angevare, L. Pedalà, U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "A 2800- μm^2 thermal-diffusivity temperature sensor with VCO-based readout in 160-nm CMOS," in *Proc. A-SSCC*, Nov. 2015, pp. 1–4.
- [14] T. Anand, K. A. A. Makinwa, and P. K. Hanumolu, "A VCO based highly digital temperature sensor with 0.034 °C/mV supply sensitivity," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2651–2663, Nov. 2016.
- [15] G. Chowdhury and A. Hassibi, "An on-chip temperature sensor with a self-discharging diode in 32-nm SOI CMOS," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 59, no. 9, pp. 568–572, Sep. 2012.
- [16] C. P. L. van Vroonhoven and K. A. A. Makinwa, "A CMOS temperature-to-digital converter with an inaccuracy of ± 0.5 °C (3σ) from -55 to 125°C," in *ISSCC Dig. Tech. Papers*, Feb. 2008, pp. 576–637.
- [17] S. M. Kashmiri, S. Xia, and K. A. A. Makinwa, "A temperature-to-digital converter based on an optimized electrothermal filter," *IEEE J. Solid-State Circuits*, vol. 44, no. 7, pp. 2026–2035, Jul. 2009.
- [18] T. Veijola and M. Andersson, "Combined electrical and thermal parameter extraction for transistor model," in *Proc. Eur. Conf. Circuit Theory Design*, Jun. 1997, pp. 754–759.
- [19] A.-J. Annema, B. Nauta, R. van Langevelde, and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 132–143, Jan. 2005.
- [20] R. Quan, U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "A 4600 μm^2 1.5 °C (3σ) 0.9kS/s thermal-diffusivity temperature sensor with VCO-based readout," in *ISSCC Dig. Tech. Papers*, Feb. 2015, pp. 488–489.
- [21] J. Robert, G. C. Temes, V. Valencic, R. Dessoulavy, and P. Deval, "A 16-bit low-voltage CMOS A/D converter," *IEEE J. Solid-State Circuits*, vol. 22, no. 2, pp. 157–163, Apr. 1987.
- [22] M. Z. Straayer and M. H. Perrott, "A 12-bit, 10-MHz bandwidth, continuous-time $\Sigma\Delta$ ADC with a 5-bit, 950-MS/s VCO-based quantizer," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 805–814, Apr. 2008.
- [23] M. H. Perrott, *CppSim System Simulator Package*, accessed on Jun. 21, 2016. [Online]. Available: <http://www.cppsim.com>
- [24] C. V. Vroonhoven and K. Makinwa, "Thermal diffusivity sensors for wide-range temperature sensing," in *Proc. IEEE SENSORS*, Oct. 2008, pp. 764–767.
- [25] X. Li, K. Maute, M. L. Dunn, and R. Yang, "Strain effects on the thermal conductivity of nanostructures," *Phys. Rev. Lett. B*, vol. 81, no. 24, p. 245318, Jun. 2010.
- [26] M.-C. Chuang, C.-L. Tai, Y.-C. Hsu, A. Roth, and E. Soenen, "A temperature sensor with a 3 sigma inaccuracy of ± 2 °C without trimming from -50 °C to 150 °C in a 16nm FinFET process," in *Proc. 41st Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2015, pp. 271–274.
- [27] M. Eberlein and I. Yahav, "A 28 nm CMOS ultra-compact thermal sensor in current-mode technique," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2016, pp. 1–2.
- [28] K. A. A. Makinwa, "Smart temperature sensors in standard CMOS," *Procedia Eng.*, vol. 5, pp. 930–939, Sep. 2010.



Uğur Sönmez (S'10–M'15) was born in Istanbul, Turkey, in 1986. He received the B.Sc. and M.Sc. degrees in electronics from Middle East Technical University, Ankara, Turkey, in 2008 and 2011, respectively. He is currently pursuing the Ph.D. degree in thermal-diffusivity-based temperature sensors with the Delft University of Technology, Delft, The Netherlands.

In 2011, he was with the Electronic Instrumentation Laboratory, Delft University of Technology. His current research interests include low-noise sensor interfaces, precision and low-power analog circuits, oversampled data converters, and time-to-digital converters.



Fabio Sebastiano (S'09–M'10) was born in Teramo, Italy, in 1981. He received the B.Sc. (*cum laude*) and M.Sc. (*cum laude*) degrees in electrical engineering from the University of Pisa, Pisa, Italy, in 2003 and 2005, respectively, the Diploma di Licenza degree from Scuola Superiore Sant'Anna, Pisa, in 2006, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011.

From 2006 to 2013, he was with NXP Semiconductors Research, Eindhoven, The Netherlands, where he focused on fully integrated CMOS frequency references, deep-submicron temperature sensors, and area-efficient interfaces for magnetic sensors. In 2013, he joined Delft University of Technology, where he is currently an Assistant Professor. He has authored or co-authored one book, seven patents, and over 30 technical publications in his research fields. His current research interests include sensor read-outs, fully-integrated frequency references, and cryogenic electronics for quantum applications.



Kofi A. A. Makinwa (M'97–SM'05–F'11) received the B.Sc. and M.Sc. degrees from Obafemi Awolowo University, Ife, Nigeria, in 1985 and 1988, respectively, the M.E.E. degree from Philips International Institute, Eindhoven, The Netherlands, in 1989, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2004.

From 1989 to 1999, he was a Research Scientist with Philips Research Laboratories, Eindhoven, where he focused on interactive displays and digital recording systems. In 1999, he joined Delft University of Technology, where he is currently an Antoni van Leeuwenhoek Professor and Head of the Microelectronics Department. He has authored or co-authored ten books, 25 patents, and over 200 technical papers in his research fields. His current research interests include the design of precision mixed-signal circuits, sigma-delta modulators, smart sensors, and sensor interfaces.

Dr. Makinwa is an Alumnus of the Young Academy of the Royal Netherlands Academy of Arts and Sciences and an Elected Member of the IEEE Solid-State Circuits Society AdCom, the Society's Governing Board. He is with the program committees of the International Solid-State Circuits Conference (ISSCC), the VLSI Symposium, the European Solid-State Circuits Conference (ESSCIRC), and the Advances in Analog Circuit Design (AACD) Workshop. For his doctoral research, he was awarded the 2005 Simon Stevin Gezel Award from the Dutch Technology Foundation. He is a co-recipient of several best paper awards, from the JSSC, ISSCC, Transducers, and ESSCIRC among others. He has also served as a Guest Editor of the IEEE JOURNAL OF SOLID-STATE CIRCUITS (JSSC) and as a Distinguished Lecturer of the IEEE Solid-State Circuits Society.