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

Fabio Sebastiano, *Student Member, IEEE,* Lucien J. Breems, *Senior Member, IEEE,* Kofi A. A. Makinwa, *Senior Member, IEEE,* Salvatore Drago, *Student Member, IEEE,* Domine M. W. Leenaerts, *Fellow, IEEE,* and Bram Nauta, *Fellow, IEEE*

Abstract

An NPN-based temperature sensor with digital output transistors has been realized in a 65-nm CMOS process. It achieves a batch-calibrated inaccuracy of $\pm$0.5 °C (3σ) and a trimmed inaccuracy of $\pm$0.2 °C (3σ) over the temperature range from -70 °C to 125 °C. This performance is obtained by the use of NPN transistors as sensing elements, the use of dynamic techniques, i.e. correlated double sampling and dynamic element matching, and a single room-temperature trim. The sensor draws 8.3 μA from a 1.2-V supply and occupies an area of 0.1 mm^2.

I. Introduction

Temperature sensors are used in a wide range of commercial applications, ranging from the control of domestic appliances and industrial machinery to environmental monitoring. Fabrication costs can be reduced by implementing the sensors in standard digital CMOS processes. This enables the co-integration of the read-out electronics, so that a digital temperature reading can be directly provided to, for instance, a microcontroller.

This work is funded by the European Commission in the Marie Curie project TRANDSSAT - 2005-020461.

S. Drago, F. Sebastiano, L. J. Breems and D. M. W. Leenaerts are with NXP Semiconductors, Eindhoven, The Netherlands, Email: fabio.sebastiano@nxp.com.

K. A. A. Makinwa is with the Electronic Instrumentation Laboratory, Delft University of Technology, Delft, The Netherlands.

B. Nauta is with the IC Design Group, CTIT Research Institute, University of Twente, Enschede, The Netherlands

An additional motivation for the development of CMOS temperature sensors in deep-submicron technologies has come from their use in the thermal management of microprocessors [1]–[4]. Although this requires an accuracy of only a few degrees centigrade, other applications are more demanding, e.g. the compensation of CMOS frequency references [5], [6] or MEMS oscillators [7].

CMOS temperature sensors with an inaccuracy of less than ± 0.1 °C over the military temperature range have been demonstrated in a mature technology (0.7-μm CMOS) [8], [9]. They are usually based on the temperature dependency of PNP transistors and achieve high accuracy by employing a single-temperature trim as well as precision circuit techniques, such as offset cancellation, dynamic element matching (DEM) and curvature correction[1].

The same sensing principle has been employed in temperature sensors in 65 nm [3] and 32 nm [10], but these only achieved inaccuracies of about 5 °C. This lack of accuracy is mainly due to the non-idealities of parasitic PNP transistors in deep-submicron technologies. Other sensing principles have been proposed for deep-submicron applications, such as the use of thermistors [1], the measurement of ring oscillator frequency [1] or MOS-transistor leakage [2]. These approaches require either multi-temperature trimming, or suffer from inaccuracies of a few degrees centigrade even over temperature ranges much narrower than the standard military or industrial temperature ranges. Sensors based on inverter delay have been proposed as good candidates for VLSI integration because of their compact layout. However, in a 0.35-μm CMOS prototype, a two-temperature trimming was necessary to achieve an inaccuracy of -0.4 °C to $+0.6$ °C over the range from 0 °C to 90 °C [11]. Furthermore, the sensor's power supply sensitivity was quite high: 33 °C/V, which is two orders of magnitude worse than that of PNP-based sensors.

This paper describes the design of a temperature sensor in 65-nm CMOS [12]. The aim was to demonstrate that accurate low-power low-voltage temperature sensors can still be designed in deep-submicron CMOS processes. Precision circuit techniques already adopted for larger feature-size processes have been employed, together with deep-submicron-specific techniques, such as the use of NPN bipolar transistors as sensing elements. In this way, a batch-calibrated inaccuracy of 0.5 °C (3σ) and a (single-temperature) trimmed inaccuracy of 0.2°C (3σ) from -70 °C to 125 °C have been achieved.

[1]Throughout the paper, *trimming* refers to the adjustment of any single sample based on the measurement of the sample itself; *batch-calibration* or *correction* (curvature correction, non-linear correction) refers to the adjustment of the samples by the same amount, equal for all the samples, which can be based on simulations, the measurement of multiple samples or of a single sample.

The sensor's principles of operation are presented in section II, while its main sources of inaccuracy and the techniques used to overcome them are described in section III. The circuit details are presented in section IV; experimental results are shown in section V and conclusions are drawn in section VI.

II. Principle of operation

The sensing principle of a bandgap (or bipolar-transistor-based) temperature sensors is depicted in Fig. 1. The sensor's core consists of a pair of matched bipolar transistors (diode-connected PNPs) biased by two currents with ratio n, to produce two temperature-dependent voltages V_{be} and ΔV_{be}. The base-emitter voltage of one transistor can be approximated as

$$V_{be} \approx \frac{kT}{q} \ln\left(\frac{I_{bias}}{I_S}\right) \tag{1}$$

where k is the Boltzmann's constant, T is the absolute temperature, q is the electron charge, I_{bias} is the bias current and I_S is the saturation current of the transistor. The temperature dependence of V_{be} is approximately linear: its extrapolated value at 0 K (V_{be0}) is close to the silicon bandgap voltage of about 1.2 V and its temperature coefficient is about -2 mV/°C [13].

The difference in base-emitter voltages ΔV_{be} can be computed from (1) as

$$\Delta V_{be} = \frac{kT}{q} \ln\left(\frac{nI_{bias}}{I_S}\right) - \frac{kT}{q} \ln\left(\frac{I_{bias}}{I_S}\right) = \frac{kT}{q} \ln n \tag{2}$$

This voltage is proportional to absolute temperature (PTAT) and is independent of process and bias conditions. Thus, ΔV_{be} is an accurate measure of absolute temperature.

ΔV_{be} can be fed to an analog-to-digital converter (ADC) to produce a digital temperature reading. The accuracy of this reading will then be limited by the accuracy of the ADC's voltage reference. In order to reduce the number of components and, consequently, the possible sources of inaccuracy, this voltage can also be generated by the same pair of bipolar transistors [13], [14]. As shown in Fig. 1, the temperature dependence of V_{be} can be compensated for by summing it with a scaled version of the PTAT voltage ΔV_{be}, as is usually done in bandgap references [15]. In this work, the appropriate scale factor is $\alpha = 18$. A PTAT digital output μ can then be generated by combining V_{be} and ΔV_{be} as follows:

$$\mu = \frac{\alpha \Delta V_{be}}{V_{bg}} = \frac{\alpha \Delta V_{be}}{V_{be} + \alpha \Delta V_{be}} \tag{3}$$

from which an output in degree Celsius can then be obtained by scaling:

$$D_{out} = A \cdot \mu + B \tag{4}$$

where $A \approx 600$ and $B \approx -273$ [16].

III. Sources of inaccuracy

A. Non-idealities of bipolar transistors

In the previous analysis, the relationship between I_{bias} and V_{be}, i.e. (1), was assumed to be exponential. This is only valid for large collector currents, i.e for $I_C \gg I_S$. A more accurate expression is:

$$I_C = I_S \left[\exp\left(\frac{qV_{be}}{kT}\right) - 1 \right] \approx I_S \exp\left(\frac{qV_{be}}{kT}\right) \tag{5}$$

where the approximation is valid for large enough collector currents I_C. CMOS temperature sensors are usually based on substrate PNPs [3], [8]–[10], [17]. As shown in Fig. 2(a), these consist of a p+ drain diffusion (emitter), an n-well (base) and the silicon substrate (collector) and are available in most CMOS processes. Since the silicon substrate is usually tied to ground, the PNP must be biased via its emitter [Fig. 3(a)]. While (1) is valid for this configuration under the approximation $I_C \approx I_{bias}$, it is possible to derive from (5)

$$V_{be} = \frac{kT}{q} \ln\left(\frac{I_E - I_B}{I_S}\right) = \frac{kT}{q} \ln\left(\frac{I_{bias}}{I_S}\frac{\beta}{\beta+1}\right) \tag{6}$$

where I_E and I_B are emitter and base currents and $\beta \triangleq \frac{I_C}{I_B}$ is the current gain of the transistor. The finite current gain and its spread affect both the curvature and the spread of V_{be}. The additional curvature can be compensated for by using standard methods for V_{be}-curvature compensation (see section III-D), but the additional spread directly impacts the sensor's accuracy. As can be understood from (6), this effect is negligible for high β but becomes increasingly significant as β decreases. For example, with $\beta = 5$, a 10% spread in current gain results in a temperature error of almost 0.1 °C over the military temperature range [16]. Though circuit techniques for finite current-gain compensation can be applied [8], device mismatch limits their effectiveness at low current gains.

The current gain of the substrate PNPs available in several CMOS processes is reported in Fig. 4. It approaches unity in deep-submicron processes, making it difficult to implement accurate temperature sensors with these devices. As an alternative, parasitic NPN transistors can be employed, which can be directly biased via their collectors. Lateral NPN transistors in

CMOS technology have been used in temperature sensors [18] but their $I_C - V_{be}$ characteristic deviates from (5) due to various extra non-idealities [16]. A better option is the vertical NPN [19], [20], which consists of an n+ drain diffusion (emitter), a p-well (base) and a deep n-well (collector), all standard features in deep-submicron processes [Fig. 2(b)]. Their only disadvantage is a higher sensitivity to packaging stress compared to vertical PNPs[2] [21].

As shown in Fig. 3(b), a vertical NPN can be biased via its collector, while the required base current can be easily provided by a feedback amplifier. The resulting base-emitter voltage will then be independent of the transistor's current gain. Moreover, the transistor's drain voltage is fixed by the feedback amplifier, making the collector current insensitive to supply voltage variations. It should also be noted that this circuit can tolerate lower supply voltages than a diode-connected PNP. With reference to Fig. 3(a), this requires a minimum supply voltage equal to the sum of V_{be} and the current source's headroom. Since V_{be} can be as high as ≈ 800 mV at the lower bound of the military temperature range (-55 °C) and a certain headroom is required to ensure current source accuracy, the minimum supply voltage can easily exceed 1.2 V. For the NPN circuit in Fig. 3(b), however, the supply voltage primarily has to accommodate the sum of the NPN's saturation voltage $V_{ce} \approx 0.3$ V $\ll V_{be}$ and of the current source's headroom. Although it must also ensure the functionality of the branch, comprising the base-emitter junction and the amplifier, that supplies I_B, no great accuracy is required of this branch. The minimum supply voltage can thus be significantly lower than in the case of a diode-connected PNP. This is a significant advantage in deep-submicron designs, which must operate at supply voltages of 1.2 V or lower.

B. ADC accuracy and quantization noise

The digital output μ in (3) can be obtained by connecting the bipolar core to the charge-balancing converter shown in Fig. 5 [22]. Here, a bias circuit generates a supply-independent current I_{bias}. Scaled copies of this current bias a pair of vertical NPNs at a n:1 collector current ratio and a third NPN with a current $n_{trim} I_{bias}$. The resulting voltages ΔV_{be} and V_{be} constitute the inputs of a 1st-order $\Sigma\Delta$ ADC. The ADC integrates $-V_{be}$ when the bitstream $bs = 1$ and integrates ΔV_{be} when $bs = 0$. Thanks to the negative feedback, the average input of the integrator is equal to zero, i.e. the integrated charge is balanced, which can be

[2]Under stress condition typical of plastic package ($\approx \pm$ 150 MPa), vertical NPNs shows a V_{be} variation of about 3 mV, which is equivalent to a temperature error of about 1 °C; those variation are 60% smaller for vertical PNPs [21].

expressed as

$$(1-\mu)\cdot\alpha\cdot V_{be}-\mu\cdot\Delta V_{be}=0 \tag{7}$$

where the bitstream average is $\mu=\langle bs\rangle$. From (7) it follows that the resulting μ satisfies (3).

In the practical implementation of the charge-balancing converter, a sensitive point is the implementation of the amplification factor α. An integer factor α is usually adopted, so that it can easily be realized by an array of α matched elements, e.g. capacitors [8]. The limit to the accuracy of α is therefore determined by the matching of these elements, requiring the use of dynamic element matching techniques that add to the complexity and area of the sensor.

Alternatively, the factor α can be realized by multiple integrations during the $bs=0$ phase. This is depicted in Fig. 6(a) for the case $\alpha=6$. The α amplifier in Fig. 5 is removed from the system of Fig. 5 and when $bs=0$, ΔV_{be} is integrated in $\alpha=6$ successive cycles. When $bs=1$, $-V_{be}$ is integrated in a single cycles. At the end of the $\alpha=6$ cycles, the comparator's output is updated. With this solution, a single element can be used to implement $\alpha=6$, but the drawback is that the conversion speed is traded for accuracy. This is because α times more cycles are used to obtain an accurate multiplication factor. The comparator is sampled only after a single integration for the $bs=1$ phase or after a series of α integrations for the $bs=0$ phase. An additional improvement in resolution can be achieved if the comparator is sampled more rapidly, e.g. after every integration, as shown in Fig. 6(b). Note that this is equivalent to multiple integrations with $\alpha=1$. The effectiveness of this approach is demonstrated by Matlab simulation of the 1$^{\text{st}}$-order $\Sigma\Delta$ converter with $\alpha=18$ and $\alpha=2$. The results are shown in Fig. 7, where the peak quantization error over the temperature range from -70 °C to 125 °C is plotted versus conversion time. In the simulation, the length of the different phases required by the circuit described in section IV has been used[3], i.e., respectively, for the $bs=1$ phase and the $bs=0$ phase, 390 μs and 100 μs for $\alpha=18$ and 70 μs and 100 μs for $\alpha=2$. The value $\alpha=2$ has been chosen since it corresponds to a simple circuit implementation (see section IV-D).

It should be noted that if $\alpha\neq 18$, then V_{bg} in (3) will no longer be temperature independent and the bitstream average μ will no longer be PTAT. A digital back-end (similar to the one

[3]With reference to the symbols used in section IV-D, for $\alpha=18$, the $bs=1$ phase is the same as in the case $\alpha=2$, while the length in the $bs=0$ phase has been assumed equal to $T_1+(\alpha-1)T_2=50\ \mu\text{s}+17\cdot 20\ \mu\text{s}=390\ \mu\text{s}$.

in [10]) is then required to compute a PTAT output, according to the relation

$$\mu_{PTAT} = \frac{\alpha_{PTAT} \cdot \mu}{\alpha + (\alpha_{PTAT} - \alpha)\mu} \tag{8}$$

where α_{PTAT} is the value required in (3) to obtain a PTAT output, and α is the value actually used in the charge-balancing converter.

It can be concluded that, for the same conversion time, using a smaller value of α results in lower quantization error, thanks to the increased granularity of the charge-balancing process. The only drawback is the need for a digital back-end to implement the non-linear correction described by (8). However, in a deep-submicron CMOS technology, this requires little extra chip area or power dissipation.

C. Process spread

Since accurate current references are not available in CMOS, I_{bias} is derived by forcing a well-defined voltage, e.g. ΔV_{be}, across a resistor. However, due to the spread of this resistor and the spread of I_S, the V_{be} of the biased transistor will still spread. As shown in [23], this spread is PTAT in nature, and can be cancelled simply by trimming the bias current used to generate V_{be}, i.e. by trimming n_{trim} in Fig. 5 [8]. In this way, a single-point trim is enough to compensate for process spread.

D. Non-linearity of V_{be}

In the previous sections, the temperature behavior of V_{be} has been considered to be linear. In practice, V_{be} shows a slight non-linearity mainly consisting of a second-order term [13]. Over the military temperature range, this can be as large as 1 °C [8].

The non-linearity in μ can be compensated for by making the temperature coefficient of the denominator of (3), i.e. V_{bg}, slightly positive [22]. This can be accomplished by increasing I_{bias} slightly compared to the value required to make V_{bg} temperature-independent. Any systematic residual non-linearity can then be compensated for by digital post-processing. A full conversion then consists of the following steps:

1) the charge-balancing converter is operated with $\alpha = 2$, as explained in section III-B;
2) the output bitstream bs is decimated to obtain

$$\mu = \frac{2\Delta V_{be}}{V_{be} + 2\Delta V_{be}} \tag{9}$$

3) a PTAT ratio μ_{PTAT} is computed:

$$\mu_{PTAT} = \frac{9 \cdot \mu}{1 + 8 \cdot \mu} \tag{10}$$

4) The residual non-linearity in μ_{PTAT} is compensated for with the help of a compensating polynomial.

IV. Circuit implementation

A block diagram of the sensor is shown in Fig. 8. The circuit design of the bias circuit generating I_{bias}, the bipolar front-end and the $\Sigma\Delta$ are described in detail in the following sections, together with the choice of the bias currents for the bipolar core.

A. Current level in the bipolar core

The bias currents of the NPN transistors in the bipolar core are constrained by several requirements, such as accuracy, noise and conversion speed. For low collector currents, the approximation $I_C \gg I_S$ used in (5) is not valid anymore and ΔV_{be} must be expressed as

$$\Delta V_{be} = \frac{kT}{q} \ln \left(\frac{\frac{nI_{bias}}{I_S} + 1}{\frac{I_{bias}}{I_S} + 1} \right) = \frac{kT}{q} \ln n + \frac{kT}{q} \ln \left(\frac{1 + \frac{I_S}{nI_{bias}}}{1 + \frac{I_S}{I_{bias}}} \right) \tag{11}$$

Thus, the bias current must be significantly larger than the saturation current in order to obtain an accurate PTAT voltage, especially at higher temperatures, since I_S increases rapidly with temperature and can reach pico-Ampere levels at 125 °C.

For large bias currents, the accuracy of ΔV_{be} is impaired by the parasitic resistances R_B and R_E in series with the emitter and the base junction respectively. In this case ΔV_{be} may be expressed as

$$\begin{aligned}
\Delta V_{be} &= \frac{kT}{q} \ln n + R_B(I_{B1} - I_{B2}) + R_E(I_{E1} - I_{E2}) && (12)\\
&= \frac{kT}{q} \ln n + \left[\frac{R_B}{\beta} + R_E \left(\frac{1}{\beta} + 1 \right) \right] (n-1) I_{bias} && (13)\\
&= \frac{kT}{q} \ln n + \left[\frac{R_B}{\beta} + R_E \left(\frac{1}{\beta} + 1 \right) \right] (n-1) I_{bias} && (14)\\
&= \frac{kT}{q} \ln n + R_S (n-1) I_{bias} && (15)
\end{aligned}$$

where $I_{B1,2}$ and $I_{E1,2}$ are the base and emitter currents of $Q_{1,2}$ and R_S is the equivalent series resistance [16]. Typical values for R_B and R_E are typically in the order of $100\ \Omega$

and $10\ \Omega$, respectively. Considering that the current gain β is commonly lower than 10 in deep-submicron processes, R_S will be in the order of some tens of Ohms, leading to a non-negligible temperature error for bias currents higher than a few hundred nano-Amperes.

The additional terms in (11) and (15) make ΔV_{be} non-PTAT. Moreover, those terms will give rise to extra spread, due to the process spread of I_S, R_S and I_{bias}. Fig. 9 shows the simulated effect of ΔV_{be} spread on the temperature reading for the NPN transistors available in the adopted technology, with the added assumption that the spread in I_{bias} is $\pm 20\%$. Since no accurate spread models for the parasitic resistances and saturation currents were available, these parameters were kept constant. The maximum allowable error (dashed line in Fig. 9) due to spread should be less than 10% of the target inaccuracy. It can be seen that several pairs of the design parameters n and I_{bias} meet those requirements. A larger n is preferable, because it implies a larger ΔV_{be} and consequently more relaxed requirements on the ADC. A larger bias current is also advantageous since it results in less noise. Based on these considerations, $n = 4$ and $I_{bias} = 50$ nA at room temperature have been chosen.

B. Bias circuit

In the bias circuit (Fig. 10), transistors Q_a and Q_b are biased by a low-voltage cascode mirror (A_1 and $M_1 - M_4$) with a 2:1 current ratio, forcing a PTAT voltage across polysilicon resistor $R_E = 180$ kΩ and making the emitter current I_E of Q_b supply-independent. The cascade of A_2 and M_{11} provides the base currents for Q_a and Q_b in a configuration similar to that shown in Fig. 3(b). The bias current $I_{bias} = I_E$ can be derived by generating and summing copies of the collector current I_C and the base current I_{Bb} of Q_b. If the current gains β of Q_a and Q_b were equal and consequently $I_{ba} = 2I_{Bb}$ held for their base currents, I_{bias} could be obtained by mirroring the drain current of M_{11} with a gain of 1/3 and adding it to a copy of I_C. However, since β is a (weak) function of collector current, a replica circuit is used to bias the matched transistor Q_c with the same collector current of Q_b and obtain an accurate copy of I_{Bb} through M_{12}. Copies of I_{Bb} and I_C (through M_{13} and M_7) are then summed at the input of a low-voltage current mirror [24].

Unlike PNP-based bias circuits [8], [9], [17], the circuit in Fig. 10 does not need low-offset amplifiers. This is because the loop comprising the base-emitter junctions of $Q_{a,b}$ and resistor R_E can be directly realized with NPNs but not with substrate PNPs. In the presented circuit, the function of the feedback amplifiers and the low-voltage current mirror is only to equalize their collector-base voltages. Thus, their offset specifications are relaxed. However, since the

base currents are relatively large ($\beta < 5$), the use of common-source buffers M_{11} and M_{12} minimizes the systematic offset of amplifiers A_2 and A_3, which otherwise would have to source these currents. The collector voltage is set to V_{CE0}, obtained by biasing R_{CE} with a copy of I_C.

A_2 and A_3 are implemented as current-mirror-loaded PMOS differential pairs with tail currents of 340 nA at room temperature and current-mirror loads. Their respective feedback loops are stabilized by Miller capacitors $C_{c1,2}$ and the associated zero-cancelling resistors $R_{c1,2}$. A_3 is a current-mirror OTA [25] with a PMOS input pair. The associated feedback loop is stabilized by Miller capacitor C_{c3}. This is kept reasonably small (1 pF), by using a low bias current (8 nA) combined with a mirror attenuation of 10 to keep the OTA's effective transconductance low. The bias currents of the amplifiers are scaled copies of I_C and are thus approximately PTAT and supply-independent.

Thanks to the use of NPNs and of the feedback loops, the circuit is able to work at low supply voltages and low temperatures. Simulation shows that, for the adopted process, the effect of supply variations on I_{bias} is less than 300 ppm/V down to a supply voltage of 1.2 V at -70 °C (for which $V_{be} > 800$ mV).

Due to the self-biasing nature of the circuit, a start-up circuit is required. The long transistor M_{14} generates a current I_{D14} in the order of few tens of nA, which is lower than the I_C at the correct operation point for any operating condition and process corner. This current is compared to I_C by the current comparator comprising M_{16}, M_{17} and M_{21}. If I_{D14} is larger than I_C, i.e. if the circuit has not yet started up, the difference $I_{D14} - I_C$ is mirrored by $M_{17} - M_{20}$ and used to start-up the circuit. The start-up current is delivered to the bases of $Q_{a,b}$, to resistor R_{CE} and as bias currents of $A_{1,2,3}$ (not shown in the schematic), which would otherwise be off because of the low I_C. The injection of $I_{start-up}$ makes the currents in all the branches increase and reach the stable operation point. When I_C becomes larger than I_{D14}, the current in M_{21} is zero and the start-up circuit is disabled.

C. Bipolar core

In the bipolar front-end (Fig. 11), transistors Q_1 and Q_2 are biased by an array of $n+1$ unit current sources ($n = 4$), whose current (50 nA) is derived from I_{bias}. The switches controlled by en_1 and en_2 can be configured to to generate a differential output $V_{\Sigma\Delta}$ equal to either ΔV_{be} or V_{be}. If en_1 (en_2) is high and en_2 (en_1) is low, the base-emitter junction of Q_2 (Q_1) is shorted and the drain current of M_{f2} (M_{f1}) is switched off to prevent any voltage drop

over the switch driven by $\overline{en_1}$ ($\overline{en_2}$). In this condition, $V_{\Sigma\Delta} = +V_{be1}$ ($V_{\Sigma\Delta} = -V_{be2}$). When both $en_{1,2}$ are high, the switches connected to the current source array are set to bias Q_1 and Q_2 either at a n:1 or at a 1:n collector current ratio, so that, respectively, either $V_{\Sigma\Delta} = \Delta V_{be}$ or $V_{\Sigma\Delta} = -\Delta V_{be}$. Because V_{be} and ΔV_{be} are not required at the same time, only two bipolar transistors are employed rather than the three used shown in Fig. 5.

When ΔV_{be} needs to be integrated, the accuracy of the 1:n current ratio and, hence, that of ΔV_{be} is guaranteed by a bitstream-controlled dynamic element matching (DEM) scheme, which is used to swap the current sources in a way that is uncorrelated with the bitstream [8]. In successive ΔV_{be}-integration cycles, a different current source is chosen from the array to provide the unit collector current, while the other $n-1$ sources provide the larger collector current. Mismatch errors in the current sources are thus averaged out without introducing in-band intermodulation products. Another source of error is the mismatch between Q_1 and Q_2, which can be expressed as mismatch of their saturation currents, respectively, I_{S1} and I_{S2}. This mismatch can cause errors when generating ΔV_{be} and can be cancelled by operating the $\Sigma\Delta$ modulator in the following way. When integrating ΔV_{be}, as explained in the following section, two phases are employed: in the first phase, $Q_{1,2}$ are biased so that $I_{C1} = nI_{C2}$ and $V_{be1} - V_{be2}$ is integrated; in the second phase, $Q_{1,2}$ are biased so that $I_{C2} = nI_{C1}$ and $-(V_{be1} - V_{be2})$ is integrated. The net integrated differential charge is then

$$
\begin{aligned}
Q_{\Delta V_{be}} &= C_a \left[\left(V_{be1}^{(1)} - V_{be2}^{(1)} \right) - \left(V_{be1}^{(2)} - V_{be2}^{(2)} \right) \right] && (16)\\
&= C_a \frac{kT}{q} \left[\left(\ln n + \ln \frac{I_{S1}}{I_{S2}} \right) - \left(-\ln n + \ln \frac{I_{S1}}{I_{S2}} \right) \right] && (17)\\
&= 2C_a \frac{kT}{q} \ln n && (18)
\end{aligned}
$$

where the superscripts (1) and (2) refers to the voltages in the first and second phase phase, respectively.

To trim the sensor at room temperature, V_{be} is adjusted, as explained in section III-C: the collector current of Q_1 or Q_2 can be coarsely adjusted via $n-1$ of the current sources, while the n-th is driven by a digital modulator to provide a fine trim [8].

The bases of $Q_{1,2}$ are loaded by the input capacitors of the $\Sigma\Delta$. Care must be taken to ensure stable operation of the loops around $Q_{1,2}$ for any bias of the collector current. Taking into consideration only one of them, the loop is comprised by three cascaded stages, Q_1, A_{f1} and M_{f1}. Miller compensation with resistive cancellation of the positive zero is introduced

around A_{f1}, so that the cascade of Q_1 and A_{f1} behaves like a two-stages Miller compensated amplifier. The gain-bandwidth product can be approximated as

$$GBW \approx \frac{g_{m1}}{2\pi C_{f1}} = \frac{I_{C1}q}{2\pi kTC_{f1}} \tag{19}$$

where g_{m1} and I_{C1} are the transconductance and collector current of Q_1. To ensure enough phase margin for the loop, the frequency of the poles associated with A_{f1} and M_{f1} must be larger than the worst-case GBW, i.e. that for the largest I_{C1}. A_{f1} (a current-mirror loaded differential pair) is then biased with a PTAT tail current derived from I_{bias} (equal to 400 nA at room temperature), so that its associated pole, proportional to its transconductance, moves to higher frequencies for higher temperatures, i.e. the conditions at which I_{C1} and consequently GBW are larger. The third pole due to the impedance and capacitance at the drain of M_{f1} is brought to high frequency by adding the diode-connected bipolar Q_3. The impedance of that node could have been lowered also by adding a diode-connected MOS transistor, but the use of a diode-connected bipolar is more advantageous than for two reasons. Firstly, Q_3 and Q_1 form a current mirror and the collector current of Q_3 tracks I_{C1}, so that the transconductance of Q_3, and thus the third pole, are larger for a higher GBW. Secondly, for a fixed current consumption, a higher transconductance can be usually achieved by a BJT rather than with a MOS. For a fixed bias current I, this is true if

$$g_{m,BJT} = \frac{I_C}{V_t} = \frac{\beta}{\beta+1}\frac{I}{V_t} > g_{m,MOS} = \frac{I}{n_{sub}V_t} \Leftrightarrow \beta > \frac{1}{n-1} \tag{20}$$

where n_{sub} is the MOS subthreshold slope factor and a MOS in weak inversion has been assumed, i.e. in the operation region with highest transconductance-to-current ratio. Since n_{sub} is typically between 1.2 and 1.6 (≈ 1.5 for the devices used in this work) [26], a BJT is more efficient for $\beta > 5$.

D. Sigma-Delta ADC

A 1st-order $\Sigma\Delta$ modulator (Fig. 11) is used to sample the voltages produced by the bipolar core. The modulator implements the charge-balancing principle described in section II, as can be understood from the example waveforms shown in Fig. 12. The modulator's switched-capacitor integrator is reset at the beginning of each temperature conversion. The opamp is based on a 2-stage Miller-compensated topology and achieves a minimum simulated gain of 93 dB (over process and temperature variations) with a PTAT bias current (3 μA at room temperature). Correlated Double Sampling (CDS) is used to reduce its offset and 1/f noise

[27]. During phase ϕ_1, the opamp is configured as a unity-gain buffer and the signal plus offset and flicker noise are sampled on input capacitors $C_{a1,2} = 2$ pF. In the second phase ϕ_2, the offset and low frequency noise are cancelled and the charge on the input capacitors is dumped on integrating capacitors $C_{b1,2}$. Since the modulator must operate at 1.2 V, the voltage swing at the output of the integrator was scaled down by choosing $C_{b1,2} = k \cdot C_{a1,2} = 4 \cdot C_{a1,2}$. Furthermore, as shown in the timing diagram in Fig. 12, when $bs = 1$, only one BJT is biased and only one base-emitter voltage $-V_{be}$ is integrated, instead of the $-2V_{be}$ of previous work [8], [9]. Since a charge proportional to $2\Delta V_{be}$ is integrated when $bs = 0$ as shown in (18), the ratio between the charge integrated for $bs = 0$ and $bs = 1$ is equal to $-2\Delta V_{be}/V_{be}$. The factor 2 results in an equivalent factor $\alpha = 2$ in the charge-balancing conversion, as mentioned in section III-B. However, this choice means that when $bs = 1$, a V_{be}-dependent common-mode voltage will also be integrated. Imbalances in the fully differential structure of the integrator, such as mismatch in the parasitic capacitances to ground at the inverting and non-inverting input of the opamp, can result in a finite common-mode-to-differential-mode charge gain, leading to error in the output. To minimize the total integrated common-mode voltage, the sign of the input common-mode voltage is alternated in successive $bs = 1$ cycles, by setting either $V_{be1} = 0$ and $V_{be2} = V_{be}$ in ϕ_1 (period A in Fig. 12), or $V_{be1} = V_{be}$ and $V_{be2} = 0$ in ϕ_2 (period B).

As shown in Fig. 12, a longer settling time is required when one input of the modulator must switch between, say, V_{be} and 0 V, when V_{be} is being integrated, than when one of the inputs must switch between, say, V_{be1} and V_{be2} when ΔV_{be} is being integrated. To minimize the conversion time, the length of each phase of the integrator are chosen equal either to T_1 when the input switches between 0 and V_{be} or to $T_1 < T_2$ when the input switches between V_{be1} and V_{be2}.

V. Experimental results

The temperature sensor (Fig. 13) was fabricated in a baseline TSMC 65-nm CMOS process, and was packaged in a ceramic DIL package. As shown in Fig. 13, the active area measures 0.1 mm^2 and it is dominated by the capacitors of the $\Sigma\Delta$'s integrator . All transistors employed in the design are thick-oxide high-threshold devices with a minimum drawn length of 0.28 μm, in order to avoid any problem due to gate leakage, which may be significant at high temperatures. In spite of the use of high-threshold device, the sensor can still operate from a 1.2-V supply, from which it draws 8.3 μA at room temperature.

The supply sensitivity is 1.2 °C/V at room temperature, which demonstrates the low-voltage capability of the proposed NPN-based sensor. The off-chip digital back-end decimates the output of the modulator and compensates for the non-linearity.

With $\alpha = 2$, the modulator's bitstream average μ is limited, varying between 0.05 and 0.18 over the temperature range from -70 °C to 125 °C. To exploit this, a sinc^2 decimation filter was instead used instead of a traditional sinc filter, as this results in less quantization error over this limited range. The digital non-linear correction described in section III-D has been applied off-line, using a 6th-order polynomial for the correction of residual non-linearities. The conversion rate of the sensor is 2.2 Sa/s (6000 bits, $T_1 = 20$ μs, $T_2 = 50$ μs) at which it obtains a quantization-noise-limited resolution of 0.03 °C. A set of devices was measured over the temperature range from -70 °C to 125 °C. After digital compensation for systematic non-linearity, the inaccuracy (Fig. 14) was 0.5 °C (3σ, 12 devices). This improved to 0.2 °C (3σ, 16 devices) after trimming at 30 °C (Fig. 15).

A summary of the sensor's performance and a comparison to the state-of-the-art for CMOS temperature sensors is reported in Table I. The sensor's untrimmed accuracy is 10 times better than previous designs in deep-submicron CMOS and both its batch-calibrated and trimmed accuracy are comparable with sensors realized in larger-feature-size processes. Furthermore, it is capable of sensing much lower temperatures, while operating from a 1.2-V supply.

VI. Conclusions

This paper describes a temperature sensor realized in a 65-nm CMOS process with a batch-calibrated inaccuracy of ± 0.5 °C (3σ) and a trimmed inaccuracy of ± 0.2 °C (3σ) from -70 °C to 125 °C. This represents a 10-fold improvement in accuracy compared to previous deep-submicron temperature sensors, and is comparable with that of state-of-the-art sensors implemented in larger-feature size processes. These advances are enabled by the use of vertical NPN transistors as sensing elements, the use of precision circuit techniques, such as dynamic element matching and dynamic offset compensation, and a single room-temperature trim. In particular, the use of NPNs, rather than the PNPs of previous work, enables low-temperature (-70 °C) sensing while operating from a low supply voltage (1.2 V).Such NPNs can be made without process modifications by exploiting the availability of deep N-well diffusions in most deep-submicron CMOS processes. This work demonstrates that accurate temperature sensors can still be designed in advanced deep-submicron CMOS processes.

REFERENCES

[1] M. S. Floyd, S. Ghiasi, T. W. Keller, K. Rajamani, F. L. Rawson, J. C. Rudbio, and M. S. Ware, "System power management support in the IBM POWER6 microprocessor," *IBM J. Res. Develop.*, vol. 51, no. 6, pp. 733 – 746, Nov. 2007.

[2] E. Saneyoshi, K. Nose, M. Kajita, and M. Mizuno, "A 1.1V 35 μm $\times$ 35 μm thermal sensor with supply voltage sensitivity of 2°C/10% -supply for thermal management on the SX-9 supercomputer," in *IEEE Symposium on VLSI Circuits Dig. Tech. Papers*, June 2008, pp. 152 – 153.

[3] D. Duarte, G. Geannopoulos, U. Mughal, K. Wong, and G. Taylor, "Temperature sensor design in a high volume manufacturing 65nm CMOS digital process," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, Sept. 2007, pp. 221 – 224.

[4] C. Poirier, R. McGowen, C. Bostak, and S. Naffziger, "Power and temperature control on a 90nm Itanium®-family processor," in *ISSCC Dig. Tech. Papers*, Feb. 2005, pp. 304 –305 Vol. 1.

[5] F. Sebastiano, L. Breems, K. Makinwa, S. Drago, D. Leenaerts, and B. Nauta, "A low-voltage mobility-based frequency reference for crystal-less ULP radios," *IEEE J. Solid-State Circuits*, vol. 44, no. 7, pp. 2002 –2009, July 2009.

[6] M. Kashmiri, M. Pertijs, and K. Makinwa, "A thermal-diffusivity-based frequency reference in standard CMOS with an absolute inaccuracy of $\pm$0.1% from -55°C to 125°C," in *ISSCC Dig. Tech. Papers*, Feb. 2010, pp. 74 – 75, 75a.

[7] D. Ruffieux, F. Krummenacher, A. Pezous, and G. Spinola-Durante, "Silicon resonator based 3.2 μW real time clock with $\pm$10 ppm frequency accuracy," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 224 – 234, Jan. 2010.

[8] M. Pertijs, K. Makinwa, and J. Huijsing, "A CMOS smart temperature sensor with a 3σ inaccuracy of $\pm$0.1 °C from -55 °C to 125 °C," *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2805 – 2815, Dec. 2005.

[9] A. Aita, M. Pertijs, K. Makinwa, and J. Huijsing, "A CMOS smart temperature sensor with a batch-calibrated inaccuracy of $\pm$0.25°C (3σ) from -70°C to 130°C," in *ISSCC Dig. Tech. Papers*, Feb. 2009, pp. 342 – 343, 343a.

[10] H. Lakdawala, Y. Li, A. Raychowdhury, G. Taylor, and K. Soumyanath, "A 1.05 V 1.6 mW 0.45 °C 3σ-resolution $\Sigma\Delta$-based temperature sensor with parasitic-resistance compensation in 32 nm CMOS," *IEEE J. Solid-State Circuits*, no. 12, pp. 3621 – 3630, Dec. 2009.

[11] P. Chen, C.-C. Chen, Y.-H. Peng, K.-M. Wang, and Y.-S. Wang, "A time-domain SAR smart temperature sensor with curvature compensation and a 3σ inaccuracy of -0.4°C ˜ +0.6°C over a 0°C to 90°C range," *IEEE J. Solid-State Circuits*, vol. 45, no. 3, pp. 600 – 609, Mar. 2010.

[12] F. Sebastiano, L. Breems, K. Makinwa, S. Drago, D. Leenaerts, and B. Nauta, "A 1.2V 10μW NPN-based temperature sensor in 65nm CMOS with an inaccuracy of 0.2°C (3σ) from -70°C to 125°C," in *ISSCC Dig. Tech. Papers*, Feb. 2009, pp. 312 – 313, 313a.

[13] G. C. Meijer, "Thermal sensors based on transistors," *Sensors and Actuators*, vol. 10, no. 1-2, pp. 103 – 125, Sept. 1986.

[14] G. Meijer, "An IC temperature transducer with an intrinsic reference," *IEEE J. Solid-State Circuits*, vol. 15, no. 3, pp. 370 – 373, June 1980.

[15] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 4th ed. Hoboken, NJ: John Wiley and Sons, 2001.

[16] M. A. P. Pertijs and J. H. Huijsing, *Precision temperature sensors in CMOS technology*. Dordrecht, The Netherlands: Springer, 2006.

[17] K. Souri, M. Kashmiri, and K. Makinwa, "A CMOS temperature sensor with an energy-efficient Zoom ADC and an inaccuracy of $\pm$0.25°C (3σ) from -40°C to 125°C," in *ISSCC Dig. Tech. Papers*, Feb. 2009, pp. 310 – 311, 311a.

[18] P. Krummenacher and H. Oguey, “Smart temperature sensor in CMOS technology,” *Sensors and Actuators A: Physical*, vol. 22, no. 1-3, pp. 636 – 638, June 1989.

[19] J. P. Kim, W. Yang, and H.-Y. Tan, “A low-power 256-Mb SDRAM with an on-chip thermometer and biased reference line sensing scheme,” *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 329 – 337, Feb. 2003.

[20] K. Szajda, C. Sodini, and H. Bowman, “A low noise, high resolution silicon temperature sensor,” *IEEE J. Solid-State Circuits*, vol. 31, no. 9, pp. 1308 –1313, Sept. 1996.

[21] J. Creemer, F. Fruett, G. Meijer, and P. French, “The piezojunction effect in silicon sensors and circuits and its relation to piezoresistance,” *IEEE Sensors J.*, vol. 1, no. 2, pp. 98 –108, Aug. 2001.

[22] G. Meijer, R. V. Gelder, V. Nooder, J. V. Drecht, and H. Kerkvliet, “A three-terminal intergrated temperature transducer with microcomputer interfacing,” *Sensors and Actuators*, vol. 18, no. 2, pp. 195 – 206, June 1989.

[23] G. Meijer, G. Wang, and F. Fruett, “Temperature sensors and voltage references implemented in CMOS technology,” *IEEE Sensors J.*, vol. 1, no. 3, pp. 225 –234, Oct. 2001.

[24] F. You, H. Embabi, J. Duque-Carrillo, and E. Sanchez-Sinencio, “An improved tail current source for low voltage applications,” *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1173 –1180, Aug. 1997.

[25] R. J. Baker, H. W. Li, and D. E. Boyce, *CMOS circuit design, layout, and simulations*. New York, NY: IEEE, 1997, p. 637.

[26] A. Pouydebasque, C. Charbuillet, R. Gwoziecki, and T. Skotnicki, “Refinement of the subthreshold slope modeling for advanced bulk CMOS devices,” *IEEE Trans. Electron Devices*, vol. 54, no. 10, pp. 2723 –2729, Oct. 2007.

[27] C. Enz and G. Temes, “Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization,” *Proceedings of the IEEE*, vol. 84, no. 11, pp. 1584 –1614, Nov. 1996.

LIST OF FIGURES

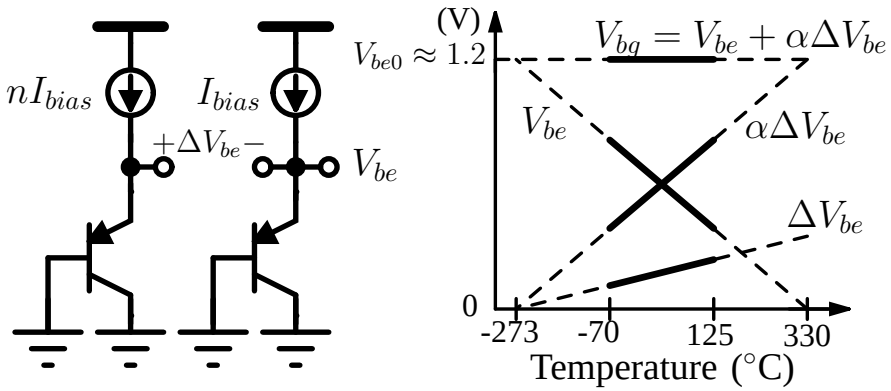

Fig. 1. Principle of operation of the temperature sensor and temperature dependance of voltages in the sensor core.

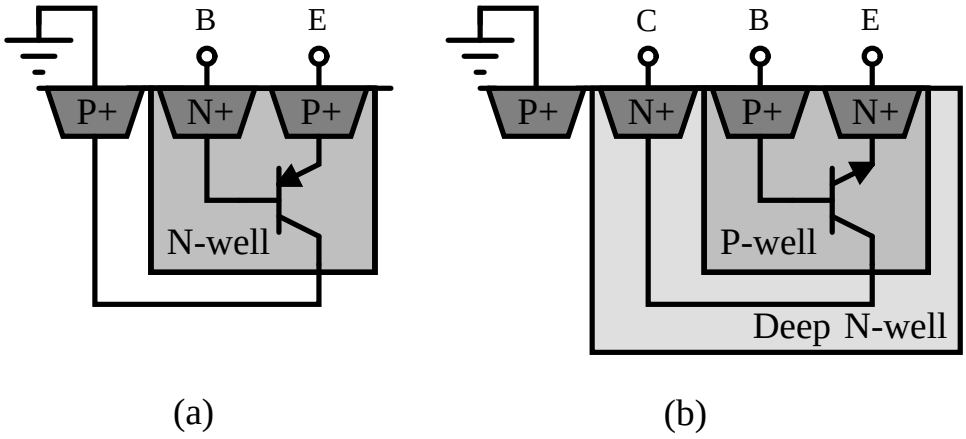

Fig. 2. Simplified cross section of (a) a substrate PNP and (b) a vertical PNP in CMOS technology.

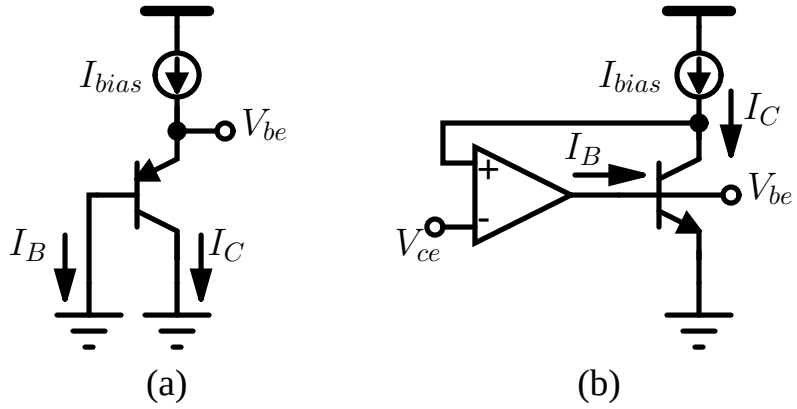

Fig. 3. Bipolar transistors configurations to generate V_{be} using (a) a substrate PNP and (b) a vertical NPN.

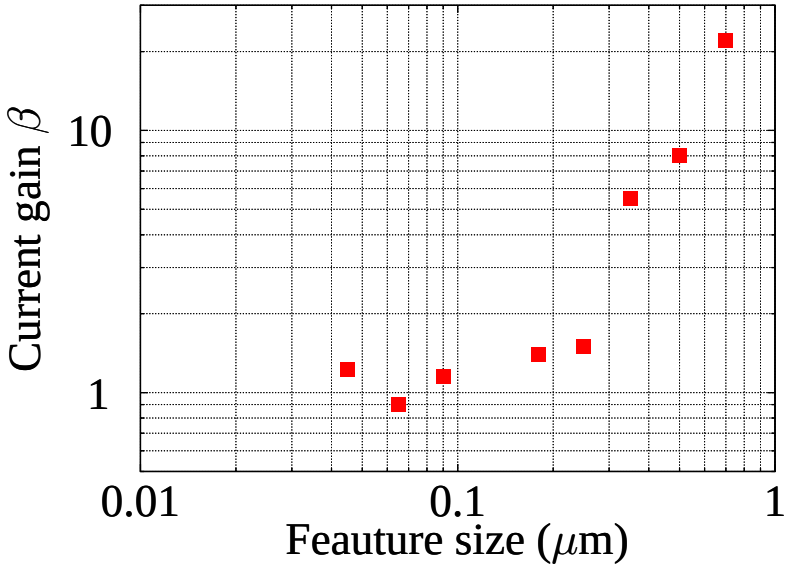

Fig. 4. Current gain β of substrate PNP transistors versus the minimum gate length for various CMOS processes (data from [16] and several design manuals).

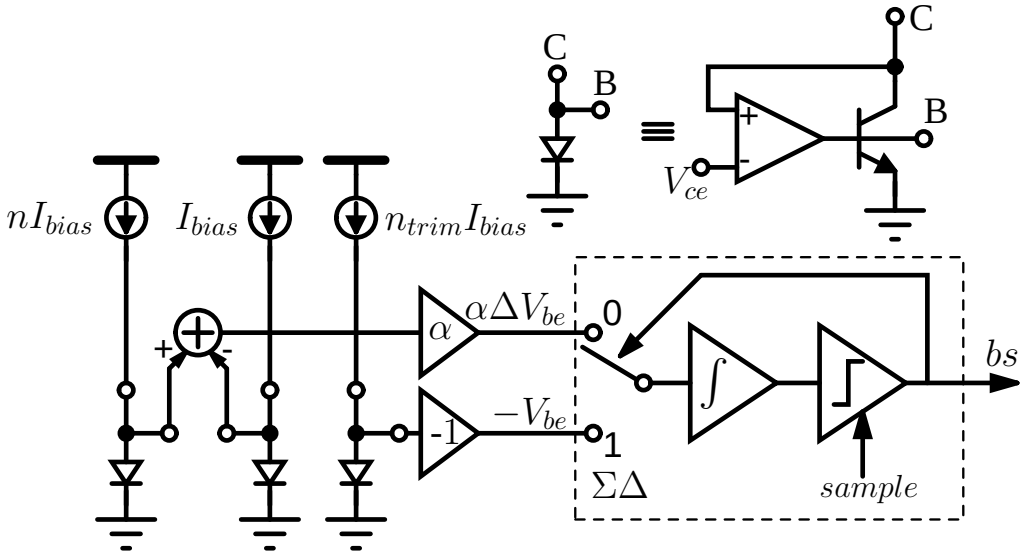

Fig. 5. Principle of operation of the charge-balancing converter.

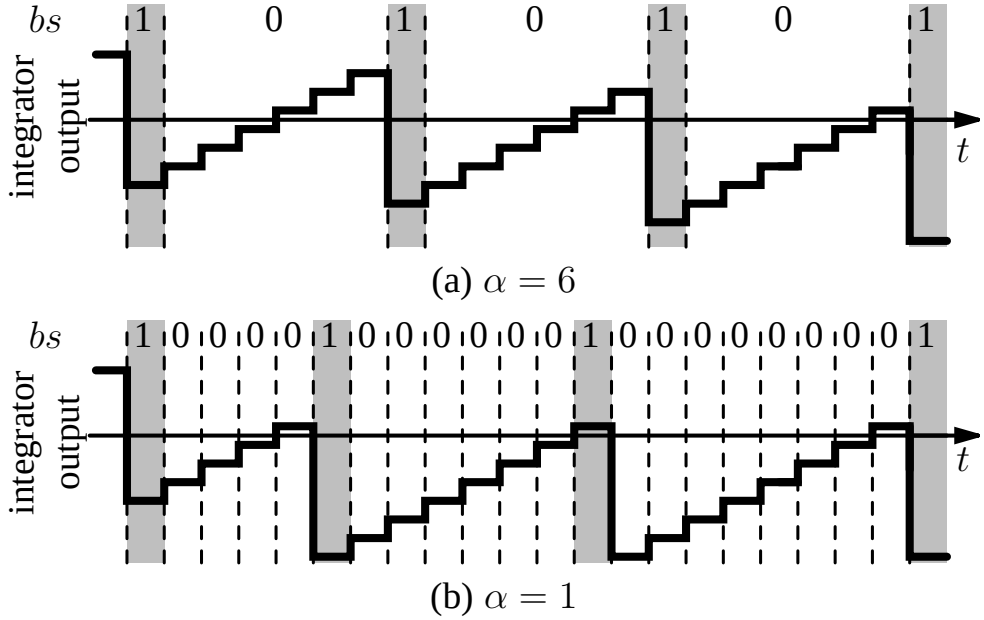

Fig. 6. Integrator output and output bitstream of a fragment of the temperature conversion of the system in Fig. 5 for different values of α; the dashed lines indicate the sampling of the comparator.

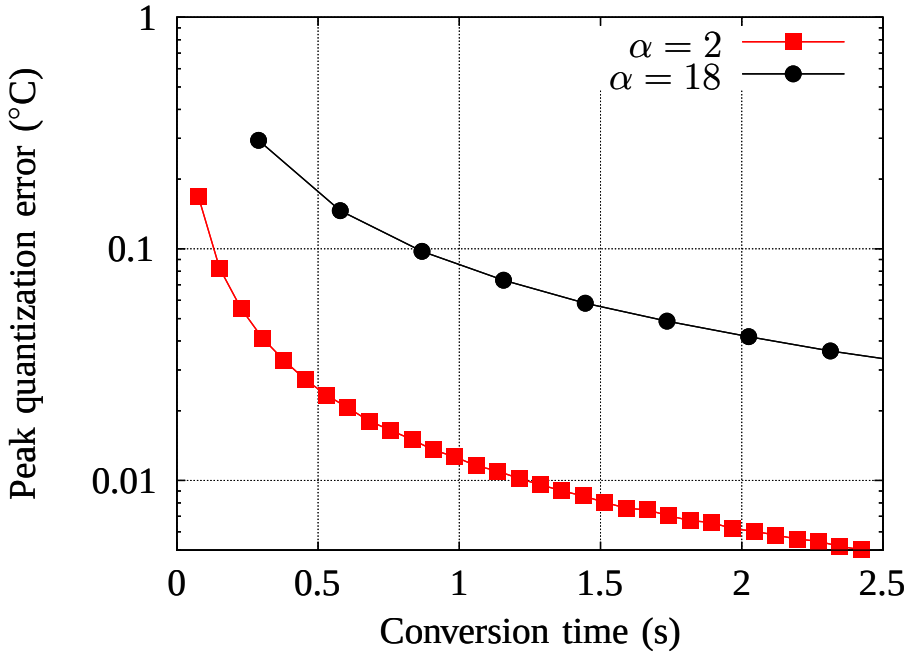

Fig. 7. Simulated peak quantization error over the temperature range from -70 °C to 125 °C versus conversion time for different values of α.

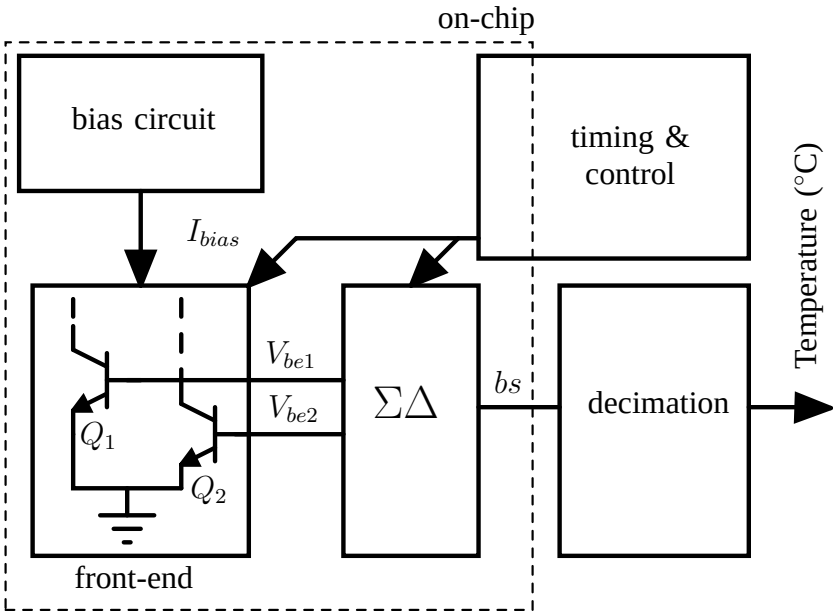

Fig. 8. Block diagram of the temperature sensor.

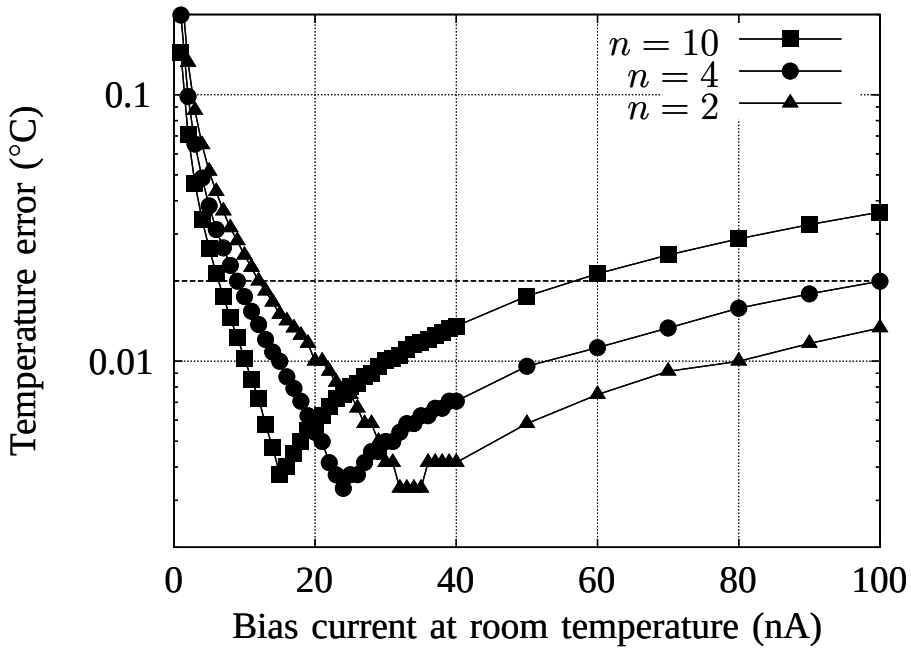

Fig. 9. Maximum temperature error over the military range due to spread in ΔV_{be} for different bias current and bias currents ratio n.

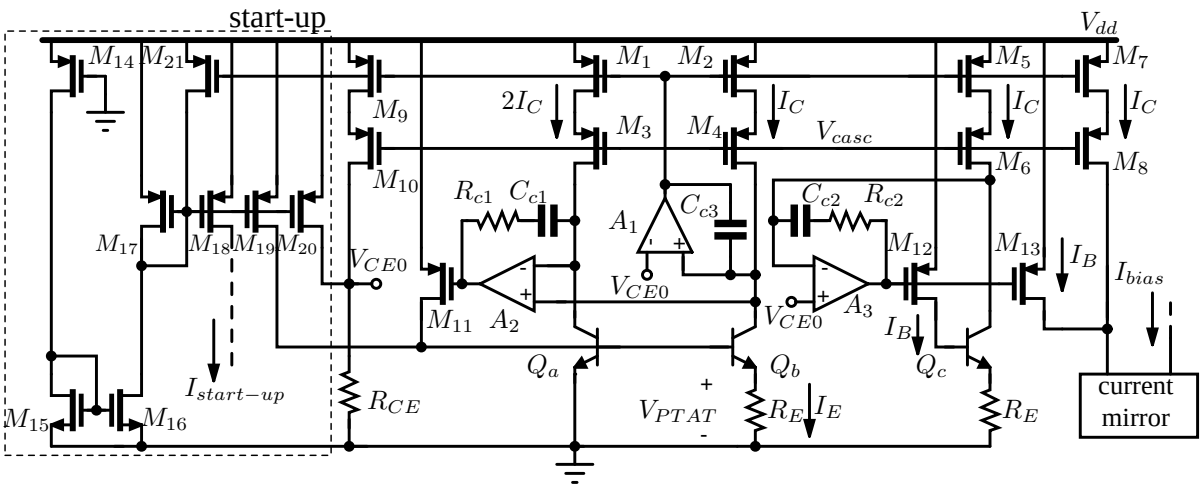

Fig. 10. Schematic of the bias circuit.

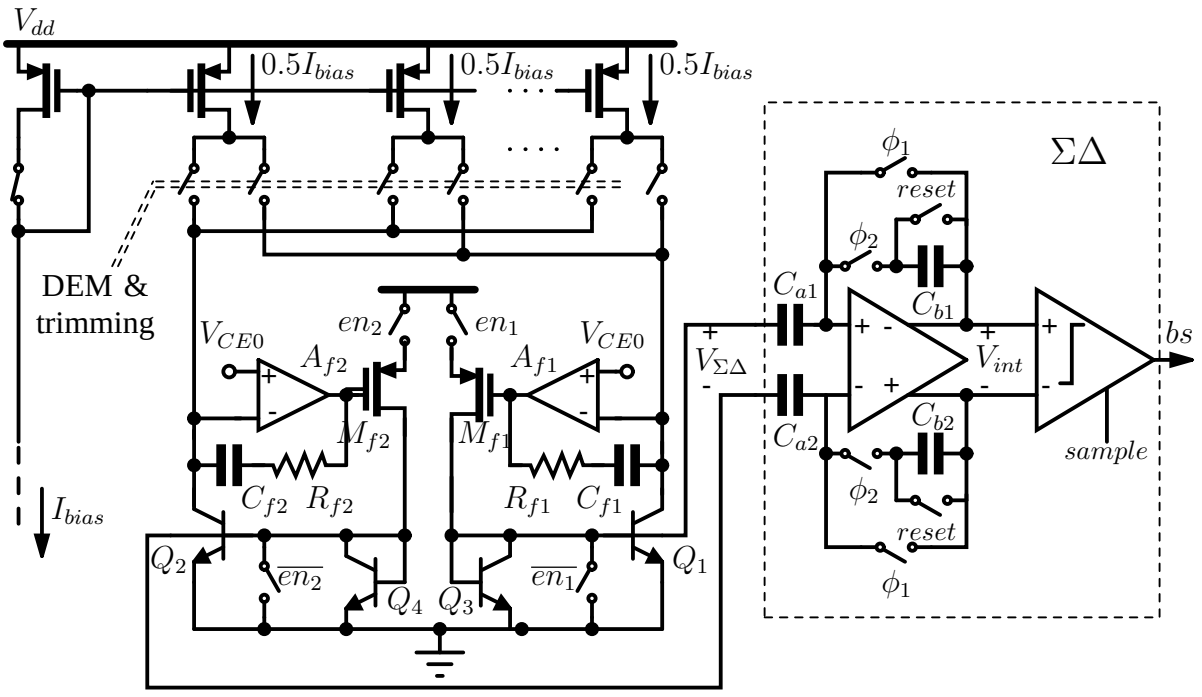

Fig. 11. Schematic of the bipolar core and of the ΣΔ ADC.

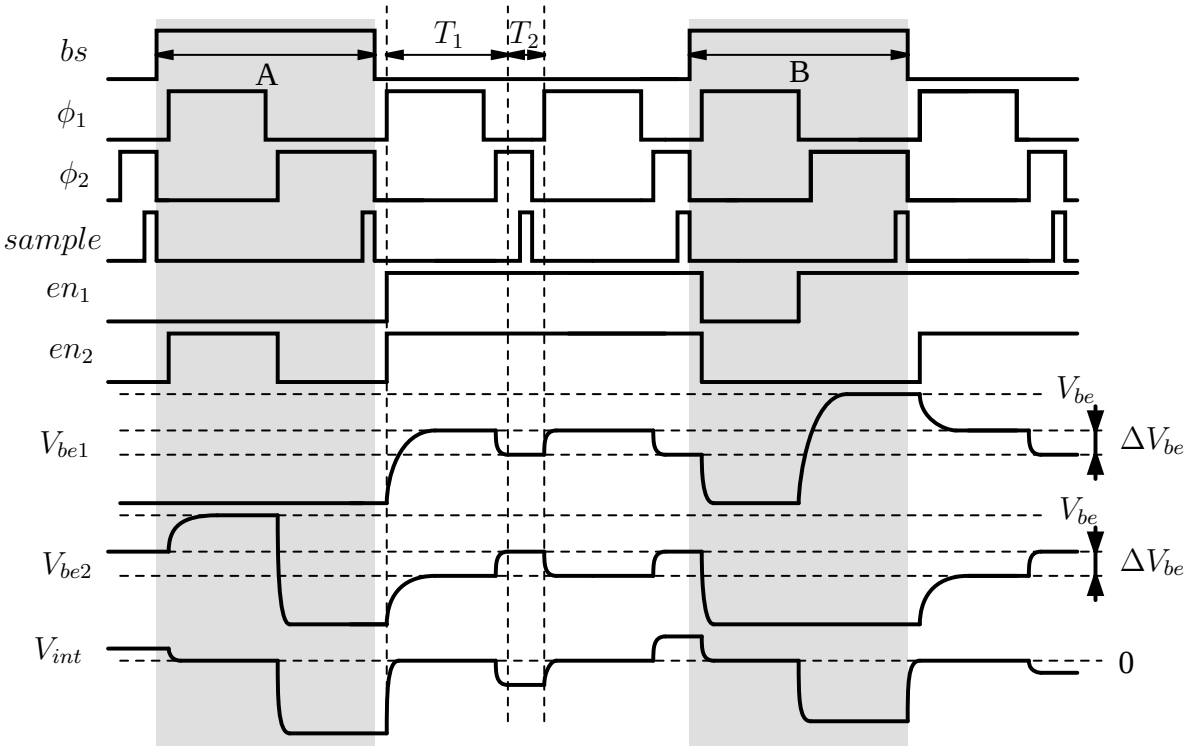

Fig. 12. Timing diagram and waveforms of a fragment of the temperature conversion; periods when bs=1 are shown in gray (A and B).

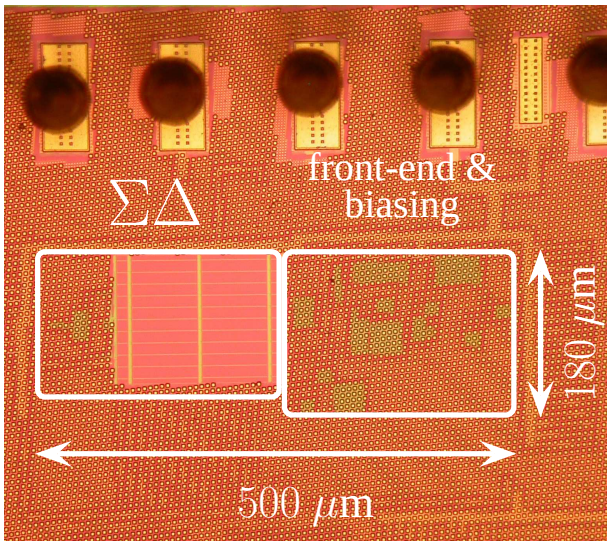

Fig. 13. Chip micrograph.

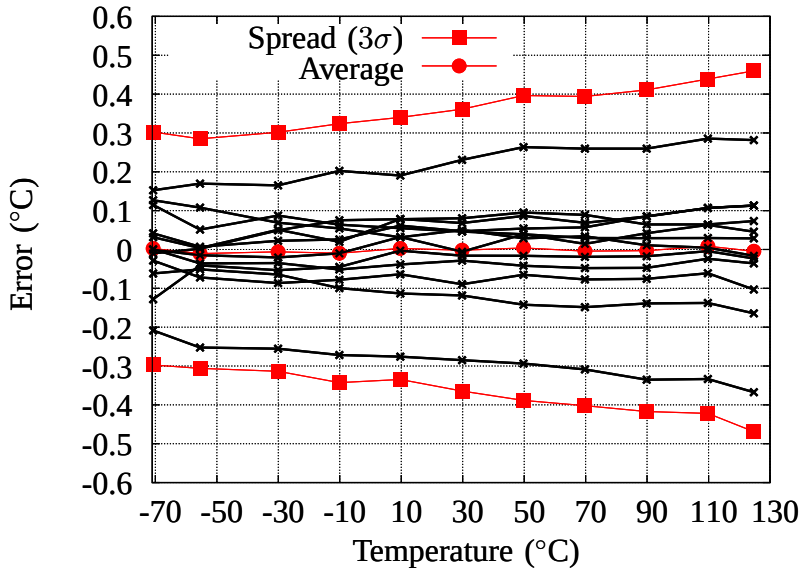

Fig. 14. Measured temperature error (with $\pm 3\sigma$ limits) of 12 samples after batch calibration.

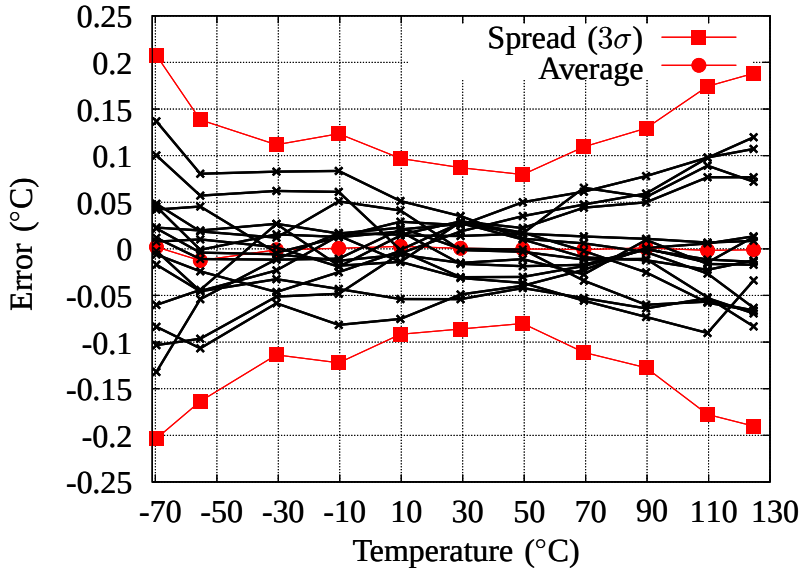

Fig. 15. Measured temperature error (with $\pm 3\sigma$ limits) of 16 samples after trimming at 30°C.

LIST OF TABLES

TABLE I

COMPARISON WITH PREVIOUSLY PUBLISHED CMOS TEMPERATURE SENSORS

Reference	This work	[10]	[17]	[9]
Technology	65 nm CMOS	32 nm CMOS	0.16 μm CMOS	0.7 μm CMOS
Chip area	0.1 mm^2	0.02 mm^2	0.26 mm^2	4.5 mm^2
Supply current	8.3 μA	1.5 mA	6 μA	25 μA
Supply voltage	1.2 - 1.3 V	1.05 V	1.8 V	2.5 - 5.5 V
Supply sensitivity	1.2 °C/V	N.A.	0.2 °C/V	0.05 °C/V
Output rate	2.2 Sa/s	1 kSa/s	10 Sa/s	10 Sa/s
Energy per conversion	4.5 μJ	1.6 μJ	0.9 μJ	12.5 μJ
Resolution	0.03°C	0.15°C (1σ)	0.018 °C (1σ)	0.025 °C (1σ)
Temperature range	-70 °C - 125 °C	-10 °C - 110 °C	-40 °C - 125 °C	-55 °C - 125 °C
Inaccuracy (untrimmed)	0.5°C (3σ)	< 5 °C	0.5 °C (3σ)	0.25 °C (3σ)
Inaccuracy (trimmed)	0.2°C (3σ)	N.A.	0.25 °C (3σ)	0.1 °C (3σ)