Highly Linear Very Compact Untrimmed On-Chip Temperature Sensor with Second and Third Order Temperature Compensation

Jun He, Chen Zhao, Sheng-Huang Lee, Karl Peterson
Dept. of Electrical and Computer Engineering
Iowa State University
Ames, IA, 50010
sirenehe@iastate.edu, zhaochen@iastate.edu, alexlsh@iastate.edu, petersok@iastate.edu

Randall Geiger, Degang Chen
Dept. of Electrical and Computer Engineering
Iowa State University
Ames, IA, 50010
rlgeiger@iastate.edu, djchen@iastate.edu

***Abstract* — This paper proposes a CMOS structure as a highly linear on-chip temperature sensor. As long as all transistors are in saturation, the output of the structure is a V_{DD} independent voltage source that linearly expresses CMOS threshold voltage, and hence is approximately linear in temperature. A new sizing strategy is introduced following a combined analytical and numerical optimization approach, which effectively removes both second and third order nonlinearities. Following this sizing approach, the sensor output voltage can be made very linear in temperature, with temperature INL (maximum temperature errors due to Vout temperature nonlinearity) within 0.05°C over the temperature range of -20~100°C. Results from corner simulations and Monte Carlo simulations demonstrate that the sensor linearity has excellent robustness over process variation and local device mismatches. With a standard two point calibration, the sensor's maximum output error can be confined within 0.15°C without any trimming. The sensor is very compact with a total active area around 200 μm² when implemented in 0.18μm process.**

I. INTRODUCTION

As the component density continues to increase in advanced CMOS technologies, power density per unit die area of VLSI chips is increasing dramatically. Reliable operations of the integrated circuit system require the prevention of excessive chip heating. Building on-chip temperature sensors to monitor the temperature at critical locations on a die is becoming an inevitable requirement. The on-chip measurement results also provide potentials to implement feedbacks from sensory data into techniques for thermal management and system performance optimization. Due to the need for many sensors throughout the die, theses on-chip sensors must be very compact. Due to the high sensitivity and nonlinear dependence of device reliability on temperature, these temperature sensors must have measurement accuracies in the sub 1°C or better range. Furthermore, these sensors must have low power consumption to avoid excessive self-heating.

This work has been sponsored by Semiconductor Research Corporation.

By far the most widely researched temperature sensors are based upon traditional proportional to absolute temperature (PTAT) principle and utilize temperature-dependent characteristics of the pn junction. Although this technique is widely used for building stand-alone temperature sensors, pnp elements and operational amplifiers are normally required to build those temperature sensors, leading to larger die sizes and high power consumptions [1]–[4]. Other authors have also focused on using the temperature dependence of CMOS threshold voltage V_T and mobility. The resultant output signals have either a pulse width proportional to the temperature or an oscillation frequency dependent dominantly upon temperature through mobility and threshold variations. In reality, the circuits obtained combine the effects of the temperature dependence of both mobility and the threshold voltage, and are not highly linear with respect to temperature [5]–[8]. The reported temperature errors range from +/-0.6°C to a few Celsius from author to author.

This paper presents a MOS temperature sensor structure that is more compact and more linear with respect to temperature. When the transistors are sized according to the proposed sizing strategy, both 2nd and 3rd order temperature nonlinearity can be compensated, and a highly linear sensor output voltage with respect to temperature can be obtained.

In section II, a V_{DD} independent circuit that can express threshold voltage is analyzed. Section III describes the new strategy on how to size the transistors to significantly reduce 2nd and 3rd order temperature nonlinearity. Design insights on how to achieve trade-offs between linearity, area and power consumption are provided. Section IV presents a circuit design example using the sizing strategy introduced in section III. Section V summarizes the work.

II. THRESHOLD EXTRACTION CIRCUIT DESIGN

The proposed circuit that is able to express threshold voltage is shown in Fig. 1. Temporarily neglecting the channel

978-1-4244-7773-9/10/$26.00 ©2010 IEEE

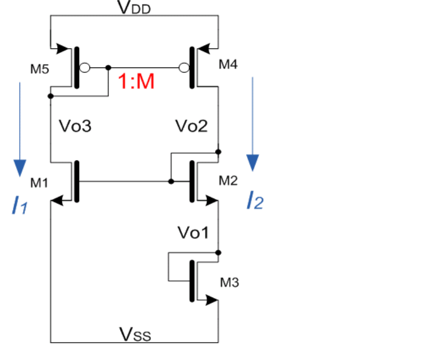

Figure 1. Schematic of proposed temperature sensor

length modulation, four equations can be written to fully describe the operation of the circuit.

$$I_1 = \frac{\mu_n C_{ox}}{2}\frac{W_1}{L_1}(V_{o2} - V_{tn1})^2 \quad (1)$$

$$I_2 = \frac{\mu_n C_{ox}}{2}\cdot\frac{W_2}{L_2}(V_{o2} - V_{o1} - V_{tn2})^2 \quad (2)$$

$$I_2 = \frac{\mu_n C_{ox}}{2}\frac{W_3}{L_3}(V_{o1} - V_{tn3})^2 \quad (3)$$

$$I_2 = M \cdot I_1 \quad (4)$$

Equations (1)–(4) comprise a set of 4 simultaneous equations in the unknown variables $\{I_{D1}, I_{D2}, V_{o1}, \text{and } V_{o2}\}$. V_{o1} and V_{o2} can be solved from the four equations.

$$V_{o1} = \frac{(V_{tn1} - V_{tn2})\cdot\sqrt{\frac{W_2/L_2}{W_3/L_3}} + V_{tn3}\cdot\left(1 - \sqrt{\frac{W_2/L_2}{M\cdot(W_3/L_3)}}\right)}{1 + \sqrt{\frac{W_2/L_2}{W_3/L_3}} - \sqrt{\frac{W_2/L_2}{M\cdot(W_1/L_1)}}} \quad (6)$$

$$V_{o2} = \frac{V_{tn1}\cdot\left(1 + \sqrt{\frac{W_2/L_2}{W_3/L_3}}\right) - V_{tn2}\cdot\sqrt{\frac{W_2/L_2}{M\cdot(W_1/L_1)}} - V_{tn3}\cdot\sqrt{\frac{W_2/L_2}{M\cdot(W_1/L_1)}}}{1 + \sqrt{\frac{W_2/L_2}{W_3/L_3}} - \sqrt{\frac{W_2/L_2}{M\cdot(W_1/L_1)}}} \quad (7)$$

where V_{tn1}, V_{tn2}, V_{tn3} are the threshold voltage for M_1, M_2 and M_3 respectively.

Assuming that all NMOS transistors have the same threshold voltage, (6) and (7) can be simplified as

$$V_{o1} = V_{tn}\frac{1 - \sqrt{\frac{W_2/L_2}{M\cdot(W_1/L_1)}}}{1 + \sqrt{\frac{W_2/L_2}{W_3/L_3}} - \sqrt{\frac{W_2/L_2}{M\cdot(W_1/L_1)}}} \quad (8)$$

$$V_{o2} = V_{tn}\frac{1 + \sqrt{\frac{(W/L)_2}{(W/L)_3}} - 2\sqrt{\frac{(W/L)_2}{M\cdot(W/L)_1}}}{1 + \sqrt{\frac{(W/L)_2}{(W/L)_3}} - \sqrt{\frac{(W/L)_2}{M\cdot(W/L)_1}}} \quad (9)$$

From (8) and (9), it can be observed that the output voltages V_{o1} and V_{o2} will have a nearly linear relationship with threshold voltage. According to threshold voltage temperature dependence model in (10), threshold voltage itself is nearly linear with respect to temperature [9],

$$V_{tn}(T) = V_{tn0} + (KT1 + KT1L/L_{eff} + KT2\cdot V_{bseff})\cdot(T/T_{NOM} - 1) \quad (10)$$

where $KT1$, $KT1L$, $KT2$ are process dependent constant, T_{NOM} is equal to 300 K, L_{eff} is the effective length approximately equal to the length L, and V_{bseff} is the effective bulk to source voltage. From (10), it can be seen that if the CMOS bulk terminal is connected to source, the temperature nonlinearity brought by bulk to source voltage V_{bs} is negligible. In the circuit in Fig.1, zero V_{bs} of M_1 and M_3 can be easily realized, while in M_2, the source and bulk cannot be easily tied together in most processes when double-well is not available. This phenomenon suggests that the PMOS counterpart of circuit in Fig.1 will potentially have better temperature linearity because each PMOS device can have its own well tie.

III. SIZING STRATEGY TO REDUCE TEMPERATURE NONLINEARITY

The channel modulation effect that has been neglected in section II will cause temperature nonlinearity in the output voltages V_{o1} and V_{o2}. This type of nonlinearity will result in several degree Celsius temperature errors. To improve the linearity, the channel modulation parameter λ is re-introduced in the analytical model in (1)–(4) to have:

$$I_1 = \frac{\mu_p C_{ox}}{2}\frac{W_5}{L_5}(V_{DD} - V_{o3} - |V_{tp}|)^2[1 + \lambda_p(V_{DD} - V_{o3})] \quad (11)$$

$$I_1 = \frac{\mu_n C_{ox}}{2}\frac{W_1}{L_1}(V_{o2} - V_{tn})^2(1 + \lambda_n V_{o3}) \quad (12)$$

$$I_2 = \frac{\mu_p C_{ox}}{2}\frac{W_4}{L_4}(V_{DD} - V_{o3} - |V_{tp}|)^2[1 + \lambda_n(V_{DD} - V_{o2})] \quad (13)$$

$$I_2 = \frac{\mu_n C_{ox}}{2}\frac{W_2}{L_2}(V_{o2} - V_{o1} - V_{tn})^2[1 + \lambda_n(V_{o2} - V_{o1})] \quad (14)$$

$$I_2 = \frac{\mu_n C_{ox}}{2}\frac{W_3}{L_3}(V_{o1} - V_{tn})^2(1 + \lambda_n\cdot V_{o2}) \quad (15)$$

The next objective is to find out d^2V_{o1}/dT^2 and d^2V_{o2}/dT^2 expressions and identify the design variables to reduce the 2nd order temperature dependence terms and compensate the quadratic temperature error. The most straightforward way to find out d^2V_{o1}/dT^2 and d^2V_{o2}/dT^2 is to directly solve V_{o1} and V_{o2} from (11)–(15). However, the highly nonlinear forms of V_{o1} and V_{o2} make it very tedious to first solve V_{o1} and V_{o2} directly and then apply differentiation. According to chain rule for implicit function differentiation theorem, we can directly apply differentiation to equations (11)–(15) with respect to temperature and then solve for first order, second order and higher order temperature derivative terms [10].

Using this approach, the first order temperature dependence terms dV_{o1}/dT and dV_{o2}/dT can be found from (16). They determine the slope of the temperature sensor transfer function.

$$
\begin{bmatrix} 0 & g_{m1} & g_{m5} \\ g_{o2}+g_{m2} & -g_{o2}-g_{o4}-g_{m2} & -g_{m4} \\ g_{o2}+g_{o3}+g_{m2}+g_{m3} & -g_{o2}-g_{m2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial V_{o1}}{\partial T} \\ \frac{\partial V_{o2}}{\partial T} \\ \frac{\partial V_{o3}}{\partial T} \end{bmatrix} = \begin{bmatrix} K_1 \\ K_2 \\ K_3 \end{bmatrix} \quad (16)
$$

In (16), $K_1 = g_{m1} \cdot \frac{\partial V_{tn1}}{\partial T} - g_{m5} \cdot \frac{\partial V_{tp}}{\partial T}$, $K_2 = g_{m4} \cdot \frac{\partial V_{tp}}{\partial T} - g_{m2} \cdot \frac{\partial V_{tn2}}{\partial T}$, $K_3 = -g_{m2} \cdot \frac{\partial V_{tn2}}{\partial T} + g_{m3} \cdot \frac{\partial V_{tn3}}{\partial T}$, g_m is transconductance, and g_o is output conductance.

Similarly, the second order temperature derivative terms can also be solved and are given in (17).

$$
\begin{bmatrix} 0 & g_{m1} & g_{m5} \\ g_{o2}+g_{m2} & -g_{o2}-g_{o4}-g_{m2} & -g_{m4} \\ g_{o2}+g_{o3}+g_{m2}+g_{m3} & -g_{o2}-g_{m2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial^2 V_{o1}}{\partial T^2} \\ \frac{\partial^2 V_{o2}}{\partial T^2} \\ \frac{\partial^2 V_{o3}}{\partial T^2} \end{bmatrix} = \begin{bmatrix} K_{21} \\ K_{22} \\ K_{23} \end{bmatrix} \quad (17)
$$

Assume that V_{o1} is the voltage of interest that we would like to linearize in the temperature domain. The 2nd temperature derivative of V_{o1} can be explicitly expressed as

$$
d^2 V_{o1} / dT^2 \approx \frac{K_{21} + K_{22}}{g_{m2}} \quad (18)
$$

where

$$
K_{21} = 2\lambda_p \cdot \mu_p C_{ox} \frac{W_5}{L_5} \cdot V_{EB5} \cdot (\frac{\partial V_{o3}}{\partial T} + \frac{\partial V_{tp}}{\partial T}) \cdot |\frac{\partial V_{o3}}{\partial T}| + \mu_p C_{ox} \frac{W_5}{L_5} (\frac{\partial V_{o3}}{\partial T} + \frac{\partial V_{tp}}{\partial T})^2
$$
$$
-2\lambda_n \mu_n C_{ox} \frac{W_1}{L_1} \cdot V_{EB1} (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{tn1}}{\partial T}) \cdot \frac{\partial V_{o3}}{\partial T}
$$
$$
K_{22} = -2\lambda_p \cdot \mu_p C_{ox} \frac{W_4}{L_4} \cdot V_{EB4} \cdot (\frac{\partial V_{o3}}{\partial T} + \frac{\partial V_{tp}}{\partial T}) \cdot |\frac{\partial V_{o2}}{\partial T}| - \mu_p C_{ox} \frac{W_4}{L_4} (\frac{\partial V_{o3}}{\partial T} + \frac{\partial V_{tp}}{\partial T})^2
$$
$$
+2\lambda_n \mu_n C_{ox} \frac{W_2}{L_2} \cdot V_{EB2} (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{tn2}}{\partial T}) \cdot (\frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{o2}}{\partial T})
$$
$$
K_{23} = -2\lambda_n \mu_n C_{ox} \frac{W_2}{L_2} \cdot V_{EB2} \cdot (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{tn2}}{\partial T}) \cdot (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{o1}}{\partial T})
$$
$$
-\mu_n C_{ox} \frac{W_2}{L_2} (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{tn2}}{\partial T})^2 [1 + \lambda_n (V_{o2} - V_{o1})]
$$
$$
+2\lambda_n \mu_n C_{ox} \frac{W_3}{L_3} \cdot V_{EB3} (\frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{tn3}}{\partial T}) \cdot \frac{\partial V_{o1}}{\partial T} + \mu_n C_{ox} \frac{W_2}{L_2} \cdot (1 + \lambda_n V_{o1}) \cdot (\frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{tn3}}{\partial T})^2
$$

where V_{EB} is the excess bias voltage of each transistor.

To reduce the quadratic term expressed by (18), the objective is to minimize the sum of K_{21} and K_{22}. There are multiple solutions to realize this objective. One solution applied in this work is:

$$
\frac{W_5}{L_5} = \frac{W_4}{L_4} \quad (19)
$$

$$
V_{EB4} \cdot |\frac{\partial V_{o3}}{\partial T}| = V_{EB5} \cdot |\frac{\partial V_{o2}}{\partial T}| \quad (20)
$$

$$
\frac{W_1}{L_1} \cdot V_{EB1} (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{tn1}}{\partial T}) \cdot \frac{\partial V_{o3}}{\partial T} \quad (21)
$$
$$
= \frac{W_2}{L_2} \cdot V_{EB2} (\frac{\partial V_{o2}}{\partial T} - \frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{tn2}}{\partial T}) \cdot (\frac{\partial V_{o1}}{\partial T} - \frac{\partial V_{o2}}{\partial T})
$$

To satisfy the conditions (19), PMOS current mirrors are chosen to have the same dimensions. dV_{o1}/dT, dV_{o2}/dT and dV_{o3}/dT in (20)–(21) are the values computed from (16). M_1, M_2 and M_3 need to be sized in a way to satisfy conditions (20) and (21) as closely as possible. In this way, the quadratic term in V_{o1} can be significantly reduced. The temperature error caused by nonlinearity can be maintained at about 1^oC level. Ideally, the higher order derivatives, such as d^3V_{o1}/dT^3 and d^4V_{o1}/dT^4 can be obtained in the same way as described above. However, the tediousness of the resultant expressions grows rapidly when moving to higher order.

To compensate 3rd order temperature nonlinear term and further reduce temperature nonlinearity, some finer size adjustment is required. First, we keep PMOS size as a constant, because it has been found that PMOS size is much less influential than the bottom three NMOS transistors in output voltage temperature non-linearity. M_1 size is critical to the current consumption of the whole circuit. Given certain power budget and voltage headroom, the size of M_1 can be approximately determined first. The lengths of M_2 an M_3 can be first fixed as 2–4 times of feature size. Therefore, two design variables $\{W_2, W_3\}$ are available for the size adjustment at first. A numerical optimization procedure to reduce temperature nonlinearity can be operated as follows:

I. Vary W_2 by a certain amount ΔW_2, such as 20% of original size, and find out the sensitivity of output temperature INL with respect to the width's change. If the INL decreases, replace the original W_2 with the new value.

II. Similarly, adjust W_3 using the approach in step I.

III. Repeat step I again and vary W_2 by a certain amount of step. This step amount change can be roughly determined according to the percentage change in the temperature INL when W_2 was varied by ΔW_2 in the previous iteration step. In general, the change in W_2 will lead to a local minimum of temperature INL error.

IV. Repeat step II to vary W_3 in the same way as varying W_2 in step III.

V. Repeat the iteration in steps III and IV, and adjust W_2 and W_3 till temperature INL does not have obvious improvement. Then consider finer adjustment in PMOS size and length of M_2 and M_3.

Using this combined analytical and numerical approach, temperature INL can be reduced to less than 0.1^oC level.

IV. A DESIGN EXAMPLE AND ITS PERFORMANCES

To demonstrate the good temperature linearity property of the circuit in Fig.1 and the effectiveness of the sizing strategy, one circuit has been designed in 1P6M 0.18um process using BSIM3v3 model. Define the temperature error caused by temperature non-linearity of the sensor's output voltage as: The maximum temperature difference between the transfer curve V_{out} versus temperature and its ending point fit line. The temperature error at typical condition is shown in Fig.2. It can be observed that after size optimization, 2nd and 3rd temperature nonlinearity term can be significantly reduced so that error caused by the temperature nonlinearity can be maintained at around 0.05^oC throughout the temperature range -20 to 100^oC.

The effect of global parameter variations from different corners is also investigated. Simulation results in Fig.3 show that the circuit has worst case temperature error at 0.15°C, which demonstrates good robustness of the design to different process corners. In the worst corner—slow NMOS slow PMOS, the threshold extraction voltage V_{o2} is increasing higher due to the larger threshold voltage. The transistor M_4 loses headroom in its V_{DS} voltage, and therefore tends to operate in the triode region and degrades the temperature linearity predicted in section II.

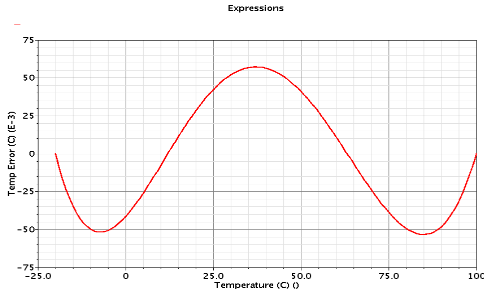

Figure 2. Temperature Error at typical condition

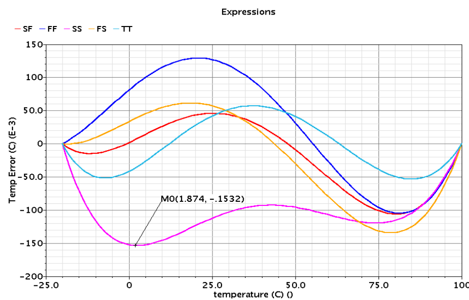

Figure 3. Temperature Error at different process corners (TT: typical; FF: fast NMOS fast PMOS; FS: fast NMOS slow PMOS; SS: slow NMOS slow PMOS; SF: slow NMOS fast PMOS)

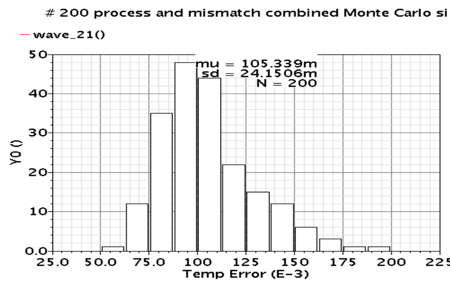

Figure 4. Temperature Error at process and mismatch combined Monte Carlo simulation

TABLE I. PERFORMANCES SUMMARY

Parameters	*Performances*
Maximum Temp Error (°C) at typical condition	0.05
Maximum Temp Error (°C) at worst corner	0.15
Maximum Temp Error (°C) at 200 # MC simulation	mu=0.105, std=0.024
Current (µA)	52
Active area (µm²)	200
Output Voltage Temp Coefficient (mV/°C)	-0.898

TABLE II. KEY CIRCUIT PARAMETERS

Process (µm)	0.18
V_{DD} (V)	1.8
Temperature range (°C)	-20 ~100
Transistor sizes	$(W/L)_1$=0.3 µ / 0.8 µ, $(W/L)_2$=2x10 µ / 0.4 µ, $(W/L)_3$=2x3.7 µ / 0.4 µ, $(W/L)_4$=4.5 µ / 0.9 µ, $(W/L)_5$=4.5 µ / 0.9 µ

In Fig.4, results from 200 times Monte Carlo simulation in 0.18µ process using BSIM3v3 model show that the circuit also has very good robustness when process variation and local device mismatches are both present. The Monte Carlo simulation models the situation when threshold voltage, mobility, transistor width and length have Gaussian distributed random variations. The mean value for the maximum temperature error is 0.105°C, and the standard deviation is 0.024°C. The main performances and design specifications have been listed in Table I. The key circuit design parameters are listed in Table II.

V. CONCLUSIONS

In this paper, a compact on-chip temperature sensor has been proposed. This structure can express the threshold voltage of CMOS transistors as outputs and achieve high temperature linearity. A sizing strategy using a combined analytical and numerical approach has been described to significantly reduce 2nd and 3rd order temperature nonlinearity. The designed circuit demonstrates temperature error at 0.05°C level and robustness to process variations and local device mismatches. The small area and high linearity makes the structure very suitable for high precision multiple sites on-chip temperature measurements.

REFERENCES

[1] Bakker, A. and Huijsing, J., "Micropower CMOS Temperature Sensor with Digital Output", IEEE J. Solid State Circuits, pp. 933-937, July 1996

[2] Bakker, A.,"CMOS Smart Temperature Sensors – An Overview", Proc. IEEE Sensors Conference, pp 1423-1427, Vol. 2, 2002

[3] M. A. P. Pertijs, K. A. A. Makinwa, and J. H. Huijsing, "A CMOS Smart Temperature Sensor with a 3σ Inaccuracy of ±1°C from -55°C to 125°C", IEEE J.Solid State Sircuits, PP. 2805-2815, Dec. 2005..

[4] M. Tuthill, "A switched-current, switched-capacitor temperature sensor in 0.6-um CMOS," IEEE J. Solid-State Circuits, vol. 33, no. 7, pp. 1117–1122, Jul. 1998.

[5] Chen, P., Chen, C., Tsai, C., and Lu, W., "A Time-to-Digital-Based CMOS Smart Temperature Sensor", IEEE Journal of Solid State Circuits, pp. 1642-1648, August 2005.

[6] B. Datta and Burleson, W., "Low-Power and Robust On-Chip Thermal Sensing Using Differential Ring Oscillators", IEEE Midwest Symposium on Circuits and Systems, pp. 29-32, August 2007

[7] Szekely, V., Marta, C., Kohari, Z. and Rencz, M., "CMOS Sensors for On-Line Thermal Monitoring of VLSI Circuits", IEEE Trans. On VLSI Systems, pp. 270-276, September 1997.

[8] Arabi, K. and Kaminska, B., "Built-In Temperature Sensors for On-Line Thermal Monitoring of Microelectronic Structures", Proc. IEEE International Conference on Computer Design (ICCD), pp. 262-465, October 1997.

[9] X. Xi, M. Dunga, J. He, W.Liu and C. Hu, BSIM4.3.0 MOSFET Model Users' Manual. Berkeley, CA: Univ. California, 2001, pp. 12-1.

[10] W.Fulks, Advanced Calculus: An introduction to analysis. John Wiley& Sons, third edition, 1978, pp.321-327.