FEASIBILITY STUDY OF ANALOG AND DIGITAL TEMPERATURE SENSORS IN NANOSCALE TECHNOLOGIES

The downscaling of CMOS technology gives rise to a myriad of nanoscale effects. At the same time, power density and thus heat generation increases. The aim of this paper is to evaluate the feasibility of both analog and digital temperature sensors in nanoscale CMOS using the Berkeley Predictive Technology Model (BPTM) for 65nm. For the oscillator-based digital sensor presented, a sensitivity of 1.86MHz/°C is achieved. The analog sensor achieves a sensitivity of 1.7mV/°C.

Introduction

As current CMOS Integrated-Circuit components have reached nanometer scales, their density is prone to give rise to higher power densities, resulting in excessive heat generation (1). This influences transistor characteristics, which can lead to reduced system performance or even false results. Furthermore, high temperatures have a direct impact on system reliability as the probability of electromigration and dielectric or junction breakdown increases.

In terms of circuit architecture, in those areas where switching is more frequent, heat generation is higher, which raises the temperature locally (hot spot), causing temperature gradients that affect the transistor behaviour (2) (3). This results in different signal delays across the chip, possibly rendering data processing lengthier than the clock phase and inducing errors.

Temperature is expected to be the limiting factor to Moore's law in the near future (1), which requires real-time adaptive circuitry suitably fed back by on-chip temperature sensors. This paper evaluates whether some current analogue and digital thermal sensing concepts are feasible in nanoscale CMOS.

Preliminaries

Temperature effects

Two major impacting parameters on MOSFET performance under temperature variations are carrier mobility and threshold voltage.

Carrier mobility In general, the mobility of holes and electrons depends on several scattering mechanisms. When considering the temperature range of interest (0-100°C), three of them are relevant (4); surface-roughness scattering, phonon scattering and Coulomb scattering. In sub-threshold operation and weak inversion, Coulomb scattering and phonon scattering dominates. As temperature increases, phonon scattering

strengthens, whereas Coulomb scattering weakens. Eventually, the doping concentration will determine the slope of the carrier mobility as a function of temperature. For high doping levels (>$10^{19}cm^{-1}$), Coulomb scattering dominates and the slope is positive. For lower doping levels, mobility is determined by phonon scattering and mobility goes down as temperature increases.

In strong inversion, mobility effects are mainly due to phonon scattering and surface-roughness scattering. The latter does not change significantly over temperature and phonon scattering dominates the carrier mobility.

Threshold voltage Several, often interrelated, parameters used to determine the threshold voltage V_{th} are dependent on temperature T. Among others, Fermi level, band gap, charge carrier concentrations are affected by temperature change (5). However, one usually approximates threshold voltage as a linear function of temperature given in equation [1]. The slope α_{vt} is expected to be in the range 1-4mV/°C (6), but it goes down for lower supply voltages (7).

$$V_{th}(T) = V_{th}(T_0) + \alpha_{Vth}(T - T_0) \qquad [1]$$

Hereby, T_0 means the temperature at which model parameters are extracted. To evaluate the sensor concepts, a predictive 65nm technology model (Berkeley Predictive Technology Model (BPTM) (8)is used, with a supply voltage V_{DD} = 900mV. The BPTM model give a slope α_{vt} of -0.3mV/°C for NMOS transistors and -0.31mV/°C for PMOS transistors. Figure 1 shows the temperature dependence of the threshold voltage of an inverter in a fanout 4 (FO4) setup.

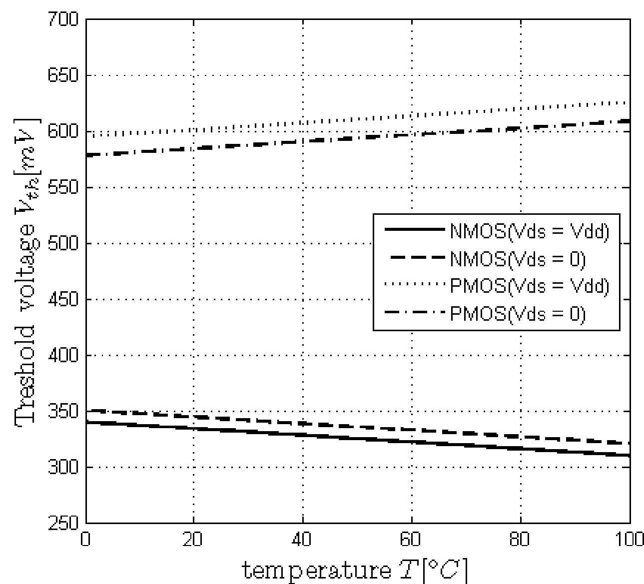

Figure 1. Threshold voltage of PMOS and NMOS transistors inside an inverter in a FO4 setup (65nm BPTM)

Consequently, the drain-source current also varies, which has high influence on the delay of logic gates. As to be seen in Figure 2, drain-source current goes down with temperature. Further, PMOS transistors show a significantly higher dependency on temperature than NMOS transistors.

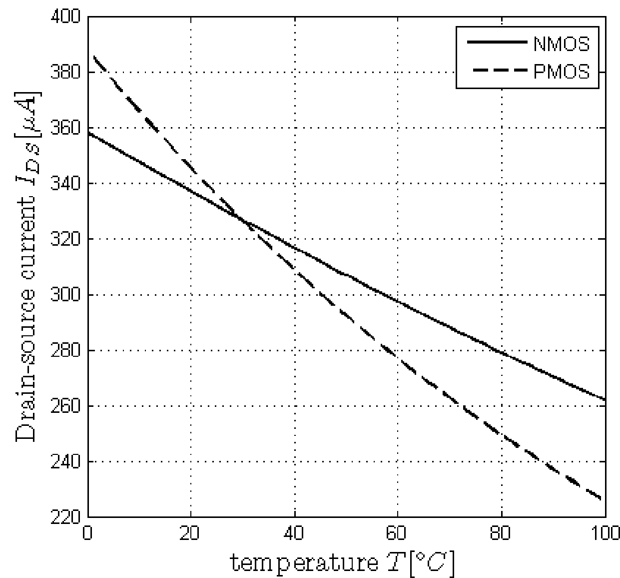

Figure 2. Drain-source current in dependence of the temperature (65nm BPTM)

Scaling effects

Smaller alignment margins, higher doping levels, closer device proximity, larger perimeter-to-area ratios and thinner dielectric layers imply higher failure rates of transistors and thus lower circuit reliability. This adds to the consequences of scaling, such as threshold voltage variations and dopant fluctuations.

The decrease in channel length implies that the source-body and drain-body depletion region occupy a larger fraction of the space-charge region controlled by the gate voltage. On top of that, when the drain voltage V_{DS} increases, the reverse biased drain-body depletion region extends further into the channel area, rendering the effective gate shorter. Eventually, also the energy barrier from source to gate will be lowered, favoring majority carriers into the channel. This results in a reduced threshold voltage (9). The difference is shown in Figure 1 for both $V_{DS} = 0$ and $V_{DS} = V_{DD} = 900mV$.

Random dopant fluctuations in the channel cause a significant variation in threshold voltage. One-micron technology had thousands of dopant atoms in the channel, whereas 65nm-technology only incorporates several hundred (3), and small atom variations become significant.

Leakage power

With the supply voltage going down the threshold voltage has to be scaled down too to meet the performance penalties. Unfortunately, such scaling increases significantly the

sub-threshold leakage current and thus the leakage power dissipation. Typically a reduction of the threshold voltage of 85mV increases the sub-threshold leakage current by a factor 10 (9).

See for further reading, extensive reviews of nanoscale effects given by Narendra (9) and Wong (10).

State of the art

Measuring temperature 'on chip' generally involves a digital output, which can be read out directly by e.g. a microprocessor. Both analog and digital temperature sensors can be summarized in a simple block scheme, as presented by Chen (11).

A smart analog temperature sensor has three basic elements, as shown in figure 3. The temperature sensor is supposed to have a high dependency on temperature, whereas the reference should be insensitive to temperature changes. An analog-to-digital converter (ADC) provides a digital output extracted from the voltage- or current difference between sensor and reference.

The digital counterpart relies on a temperature-to-time generator in the form of a delay line or ring oscillator with a high temperature dependency. The temperature information is captured in a pulse-width modulated or frequency modulated signal and converted into a digital representation of temperature by e.g. a counter. This is schematically drawn in Figure 4.

Ring oscillator

The implemented temperature-to-time generator bases on a ring oscillator as this element has, compared to a delay line, lower performance constraints. The next step is the indentification of the recommended gate types for this ring oscillator.

In a gate library corresponding to a given technology, standard cells as inverters, NANDs, NORs et cetera are available. Mostly, the width of NMOS transistors (W_n) and the width of PMOS transistors (W_p) are optimized for equal rise and fall slope of the output. For an inverter, the ratio W_p/W_n is close to 2, to make up for the lower mobility of carriers in PMOS transistors compared to NMOS transistors, at the doping levels practiced in sub-micron technologies.

The use of these standard cells leaves little room for maximization of delay, as desired for a ring oscillator. The design of custom gates to maximize the circuit delay is therefore considered. In the end, the goal is to maximize the first-order frequency variation to temperature, $\Delta F/\Delta T$. As this value increases, the requirements to the subsequent counter become easier to meet.

To compare the logic gates, the temperature range is set to 0-100°C with a maximum oscillating frequency F = 500MHz. Furthermore, transistors widths are restricted to have maximum width of 2µm to prevent exorbitant widths. To design custom gates which lead to higher circuit delay, both scaling the width of the transistors and the ratio W_p/W_n are investigated.

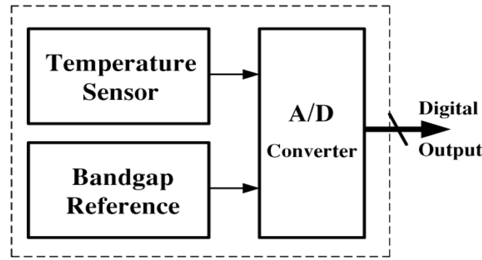

Figure 3. Block scheme of a smart analog temperature sensor (10)

The most straightforward ring oscillator that meets the requirements above is a chain of inverters with standard threshold voltage. Scaling the width of all transistors while keeping ratio W_p/W_n equal to 2 shows a minimum delay around $W_n = 400$nm and $W_p = 800$nm. To maximize delay, transistors should be either wider or narrower. For both power and area, the gates should thus be as small as possible, $W_n = 65$nm and $W_p = 130$nm. Using 63 cascaded inverters, the oscillator meets the maximum frequency specification with minimum amount of gates and reaches a sensitivity of $\Delta F/\Delta T = 1.761$MHz/°C. This and other standard gates are summarized in Table 1. In a subsequent step the influence of transistors with high- (*hvt*) and low- (*lvt*) threshold voltage was analyzed. The models were taken from a modified version of the BPTM technology (**Error! Reference source not found.**2). Of the standard gates in Table 1, the NOR3 and NAND-NOR2 have the highest sensitivity. The first uses less power, the latter uses less area.

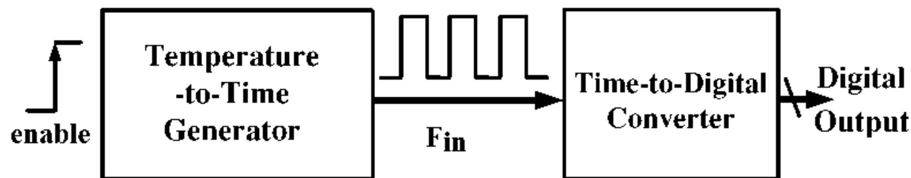

Figure 4. Block scheme of a smart digital temperature sensor (10)

When designing custom gates, one has the freedom to vary the dimensions of the transistor. In this paper the influence of ratio W_p/W_n is evaluated. In normal use of an inverter, this ratio is close to 2 for equal output slopes. The delay of an inverter chain with a range of W_p from 65nm to 2um is simulated, keeping W_n fixed at 500nm. This gives minimum delay for a ratio of 1.3. For $W_p << W_n$ and $W_p >> W_n$, the delay of the inverter chain increases drastically. Another suggestion is to alternate the ratio in such a way that a small PMOS transistor has to charge the input capacitance of a big NMOS transistor and vice versa. This cascade of inverters with alternating ratios is named 'alternating ratio' in Table 1. Table 1 shows that the inverter chain with $W_p << W_n$ has a very high sensitivity, but at the cost of area. The alternating ratio chain has a high sensitivity as well, but remarks have to be made. The signal coming from this chain is a pulse with a width of about 1ns, while withdrawing almost 1mA from the power supply.

Table 1 Ring oscillator properties

		$\Delta F / \Delta T$ [MHz/°C]	**I_{dd}, max** [uA]	**power** [uW]	**Total gate area** [pm^2]
Standard gates	INV	1,761	194,7	27,5	717
	INV_lvt	1,726	207,6	33,1	1081
	INV_hvt	1,811	185,1	24,2	808
	NAND2	1,77	265,8	58,4	2089
	NAND3	1,784	204,3	32,1	563
	NOR2	1,82	245,5	34,6	1223
	NOR3	1,845	271,3	36,6	1450
	NAND-NOR2	1,857	263,4	43,2	1223
	NAND-NOR3	1,816	197	33,7	1588
Custom gates	Alternating ratio	1,864	985,8	28,1	1196
	INV(Wp>>Wn)	1,369	283	79,2	3057
	INV(Wp<<Wn)	1,933	368	57,1	1994

Analog sensor

The goal of this section is to analyze the analog sensor proposed by Sasaki (13) in a 65nm technology (see figure 5). During operation, the two branches are biased by a PMOS current mirror. An error amplifier adjusts this bias current such that V_{out1} equals V_{out2}. Hence, the sensor settles at a point where both voltage and current are exactly the same for both branches. This introduces an operating point that shifts with temperature, as depicted in Figure 6. Firstly, the proposed sensor is implemented in the 65nm BPTM technologywith the ratios suggested by Sasaki (see figure 5), using transistors with standard threshold voltage and a supply voltage of 1V. The sensitivity achieved with these ratios is $\Delta V/\Delta T$ = 1.5mV/°C.

Variation of parameters shows room for improvement. First of all, the supply voltage is reduced to 900mV, as with the digital design. Maximizing the widths of M2 (W_2) and M3 (W_3) has a positive influence on linearity, as V_{out2} goes up and intersects with the more linear part of V_{out1}. Variation of both lengths L_2 and L_3 changes the slope of V_{out2} in the origin. When both are set to 94nm, this slope is the highest. For any other value, the slope at the origin will decrease. Possible explanation for this might be that short channel effects are no longer negligible for gate lengths smaller than 94nm. Widening of transistors 0 and 1 implies scaling the current through both transistors and increasing power while losing linearity. Adjusting L_0 and L_1 only affects V_{out1}. This adjustment is also closely related to bias voltage V_{gs0}. In all cases, V_{gs0} should not be below threshold voltage to keep transistor M_0 in saturation. Maximizing L_0 and L_1 ensures this. By altering V_{gs0}, sensitivity and linearity can be interchanged.

After this rough optimization, a sensitivity of 1.7mV/°C is achieved. The actual parameters are enlisted in Table 2. The output is plotted for several temperatures in figure 6.

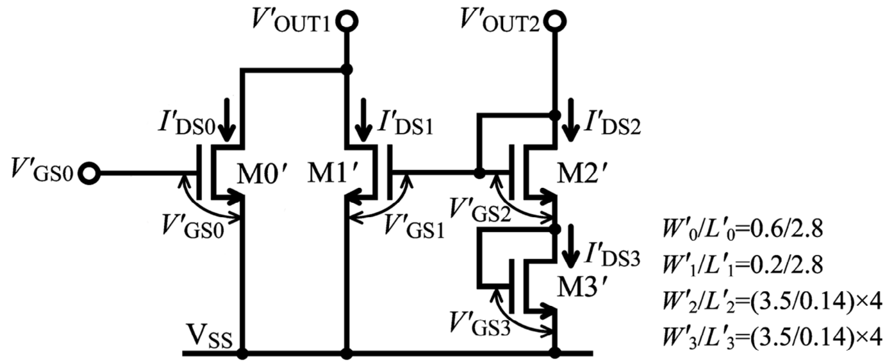

Figure 5. Four transistor analog sensor (12)

Table 2 Sensor parameters

W_0/L_0 = 0.75/4

W_1/L_1 = 0.25/4

W_2/L_2 = (4/0.094)x4

W_3/L_3 = (4/0.094)x4

V_{GS0} = 650mV

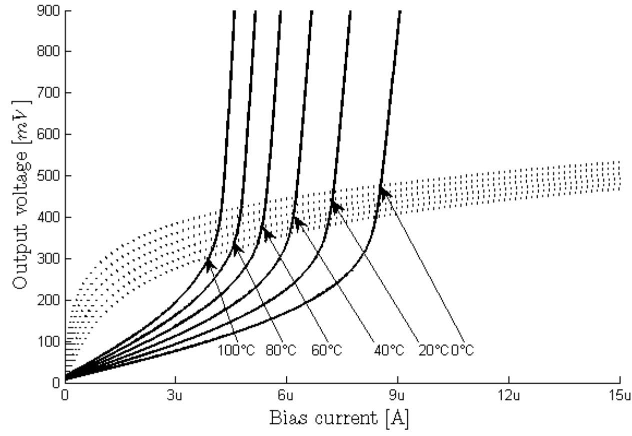

Figure 6 Operating points (0-100°C)

Comparison

The most important parameters to be compared are performance, power consumption and footprint. In this paper, performance is mainly characterized by sensitivity. Further research is necessary to take into account the linearity of the sensors, total effort including peripherals and sensitivity to parameter variations inherent to nanoscale CMOS. The achieved sensitivities impose requirements to the peripherals, such as a frequency divider and/or counter in the digital case. Equivalently, a bias voltage source, error amplifier and ADC are peripherals to be considered in the analog case. These peripherals have a major influence on both power and area.

Compared to recent work, the simulated digital sensor reaches sensitivities greater than 1.8MHz/°C, where recent work in 180nm CMOS reports 1MHz/°C (14). The analog sensor that was implemented, obtained a sensitivity of 1.7mV/°C compared to 1.8mV/°C for the original design by Sasaki (13) in 90nm CMOS. Both sensor architectures can be almost unaltered employed in 65nm technology, suggesting even increased performance.

Conclusion

Scaling down CMOS technology implies both increase of heat generation and higher parameter spread. The aim of this paper was to evaluate if temperature sensors as used in today's integrated circuits still fulfil the needs regarding sensitivity. Two sensor concepts have been implemented using the Berkeley Predictive Technology Model for 65nm with a supply voltage of 900mV. The digital sensor has a sensitivity of 1.8MHz/°C, the analog sensor obtains 1.7mV/°C. Comparing the sensitivities to recent work in deep submicron technology shows that both concepts are viable in 65nm CMOS technology.

References

1. G. E. Moore, *Solid-State Circuits Conference,* p.20- 23, **1 (2003).**
2. M. Riley, J. Warnock, and D. Wendel, *IBM Journal of Research and Development*, **51**, 5 (2007).
3. S. Borkar, *IEEE Micro*, pp. 10-16, **25**, 6 (2005).
4. S.C. Sun, *IEEE Journal of Solid-State Circuits*, **15**, 4 (1980).
5. S. Middelhoek, *Silicon sensors* (2000)
6. I. M. Filanovsky, *IEEE Transactions on Circuits & Systems*, **48**, (2001).
7. C. Park, *IEDM 95*, pp. 71-74 (1995).
8. M.V. Dunga, *BSIM4.6.0 MOSFET Model* (2006).
9. S.G. Narendra, *ACM Journal on Emerging Technologies in Computing Systems*, pp. 7-49, **1**, 1 (2005).
10. H.P. Wong, *Proc. of the IEEE*, **87**, 4 (1999).
11. C. Chen, *Measurement Science & Technology*, pp. 840-846 (2006).
12. [*omitted for blind review*]
13. M. Sasaki, *IEEE Transactions on Semiconductor Manufacturing*, **21** (2008).
14. N. Jandhyala, *9th International Symposium on Quality Electronic Design*, (2008).