A Dissertation Report on

Design of CMOS Temperature Sensor for Dynamic Voltage and Frequency Scaling Applications

Submitted By
Harshit Goel
(Enrollment. No. 17534001)

Under the guidance of
Dr. Bishnu Prasad Das
(Assistant Professor)

Department of Electronics and Communication Engineering

Indian Institute of Technology Roorkee,
Roorkee 247 667, Uttarakhand (INDIA)

May 2019

CANDIDATE'S DECLARATION

I declare that the work presented in this report with title "**Design of CMOS Temperature Sensor for Dynamic Voltage and Frequency Scaling Applications** " towards the fulfillment of the requirement for the award of the degree of Masters of Technology in Microelectronics & VLSI submitted in the Dept. of Electronics & Communication Engineering, Indian Institute of Technology, Roorkee, India is an authentic record of my own work carried out during the period from January 2018 to May 2019 under the supervision of **Prof. Bishnu Prasad Das**, Assistant Professor, Dept. of ECE, IIT Roorkee. The content of this report has not been submitted by me for the award of any other degree of this or any other institute.

DATE: .....

SIGNED: .....
(HARSHIT GOEL)
17534001
MICROELECTRONICS & VLSI

CERTIFICATE

This is to certify that the statement made by the candidate is correct to the best of my knowledge and belief.

DATE: .....

SIGNED: .....
(PROF. BISHNU P. DAS)
ASSISTANT PROFESSOR
DEPT. OF ECE, IIT ROORKEE

Acknowledgment

First and foremost, I would like to express my sincere gratitude towards my guide **Prof. Bishnu Prasad Das, Assistant Professor, Dept. of Electronics and Communication Engineering, IIT Roorkee** for his ideal guidance throughout the period. I want to thank him for the insightful discussion and constructive criticisms which certainly enhanced my knowledge as well as improved my skills. His constant encouragement, support, and motivation were key to overcome all the difficult and struggling phases.

Also, I would like to thank **Poorvi Jain**, Research Scholar, Microelectronics & VLSI, Dept. of Electronics and Communication Engineering, IIT Roorkee for guiding me whenever I faced any difficulty.

I would also thank the **Department of Electronics & Communication Engineering, IIT Roorkee** for providing me the lab and other resources for the project work.

Harshit Goel
Microelectronics & VLSI

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Motivation

Starting from 2300 transistors on a chip in 1971 by INTEL to 192 billion transistors on a single chip by AMD in 2017, the technology has grown faster than expected and so did the leakage currents and power consumption on these chips. These modern systems tend to have a behavior similar to positive feedback i.e. leakage currents increase heat dissipation, increasing the temperature of the chip which in turn exponentially increases leakage current. In a recent report it was reported by intel [1], that for every 15 °C rise in temperature delay will be increased by approximately 10% - 15%.

On-chip temperature monitoring has become inevitable in the current scenario when chips are running at exceptionally high frequencies dissipating lots of heat to chip vicinity, which creates performance and reliability issues in the microprocessor. In DRAMs, the self-refresh rate is dependent upon current die temperature. On-chip temperature sensors provide critical feedback to the dynamic voltage and frequency scaling blocks that are implemented to prevent excessive chip heating which can destroy the device or reduce the expected lifetime. Multiple small temperature sensors are distributed throughout the chip at expected hotspots to monitor on die temperature, in applications where excessive heating of chip is of concern.

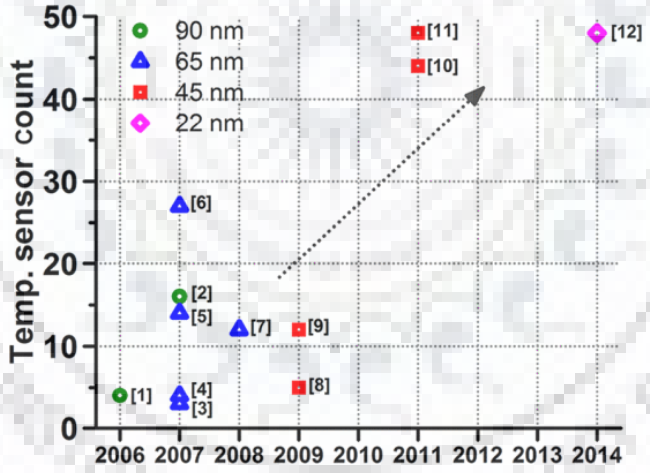


Fig. 1.1 Recent trend showing the number of on-chip temperature sensors in SoC. [3]

There are three basic requirements for designing a temperature sensor:

1. Sensors need to be small and compact. The increasing level of integration on a chip increases the number of hotspots in the chip, also the location of these hotspots cannot be predicted so as to incorporate as many as sensors in the chip.
2. The calibration cost of sensors should be low while their accuracy should be considerably high.
3. The sensors should work properly at low supply voltages (V_{DD}) and should be immune to supply voltage variations. In most of the digital circuits, these sensors are powered by local(logic) supply voltage which is very noisy due to continuous switching.

1.2 Literature Review

There are several methods by which these temperature sensors can be realized. Different authors used different approaches depending upon their requirements of the area, power consumption, sensitivity, and output [2] - [4]. There are only two device parameters which are extensively dependent on temperature i.e. threshold voltage (V_{TH}) and mobility (μ). Temperature information is converted in the form of voltage, frequency or current by exploiting the V_{TH} and μ or sometimes thermal voltage (V_T) in a circuit. In [2], a wilder architecture based temperature sensor is presented which uses V_{TH} as temperature dependent parameter. The temperature sensor shows a sensitivity of -1.12 mV/ °C. The maximum INL temperature error over a temperature range of 120 °C is about 0.4 °C. The threshold voltage change with temperature is governed by the equation of V_{TH} in BSIM model which is very linear. The very small non-linearity in the simulated results are because of the finite output impedance of the MOS transistor due to which threshold voltage can be expressed at output solely.

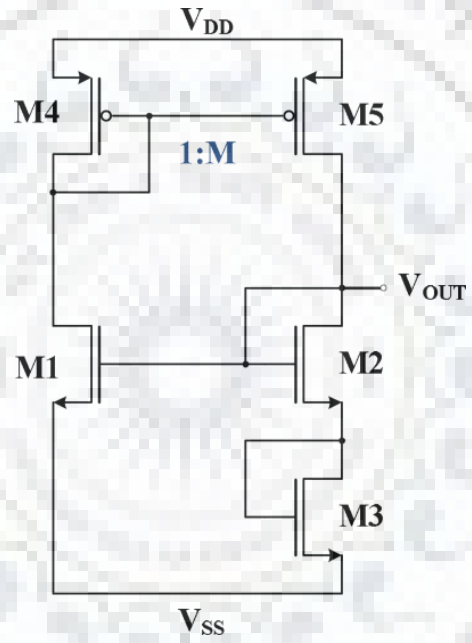


Fig. 1.2 Schematic of Inverse welder based temperature sensor circuit. [2]

The circuit diagram is shown in Fig 1.2, V_{OUT} is found to be proportional to the threshold voltage of transistor M1 and is given by the equation (1). It is quite apparent from this expression that V_{OUT} depends linearly on threshold voltage(V_{TH}) and is independent of supply voltage(V_{DD}).

$$V_{out} = \frac{V_{tn1} \cdot \left(1 + \sqrt{\frac{\left(\frac{W}{L}\right)_2}{\left(\frac{W}{L}\right)_3}}\right) - V_{tn2} \cdot \sqrt{\frac{\left(\frac{W}{L}\right)_2}{M \cdot \left(\frac{W}{L}\right)_1}} - V_{tn3} \cdot \sqrt{\frac{\left(\frac{W}{L}\right)_2}{M \cdot \left(\frac{W}{L}\right)_1}}}{1 + \sqrt{\frac{\left(\frac{W}{L}\right)_2}{\left(\frac{W}{L}\right)_3}} - \sqrt{\frac{\left(\frac{W}{L}\right)_2}{M \cdot \left(\frac{W}{L}\right)_1}}} \quad (1)$$

These models show V_{DD} independency only when V_{DD} is high and all the transistors are working in saturation mode hence area and power consumption of these architectures is quite high. This design acquires an area of 328.6 $\mu m^2$and the total power consumption is 103 µW at a supply voltage of 1.8 V. In order to eliminate an undesired stable operating point a start-up circuit is required in this design to. In [3] temperature sensors based on V_{DD}- compensated proportional to absolute temperature (PTAT) and complementary to absolute temperature (CTAT) are proposed and shown in Fig 1.3.

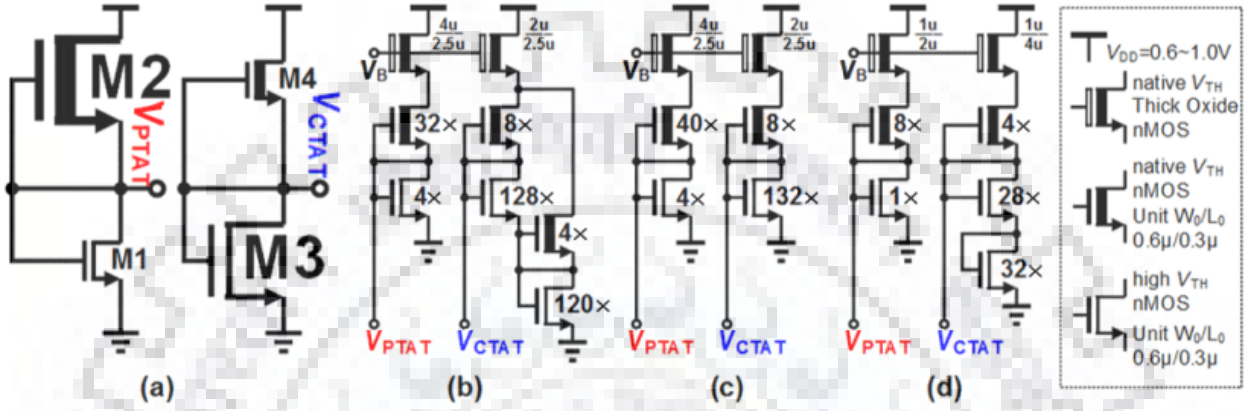


Fig. 1.3 (a) Basic structure showing front end design to find V_{PTAT} and V_{CTAT}. The implemented structures of the (b) accuracy-optimized, (c) accuracy-area-balanced, and (d) area-optimized front ends. [3]

In the circuit of Fig. 1.3 (a), M2 and M4 are NMOS transistors with nearly zero V_{TH} so that they can be activated with gate and source shorted. M1 and M3 are thin-oxide, High- V_{TH} NMOS transistors that are configured as diodes. M2 is sized larger than M1, and M4 smaller than M3, to generate V_{PTAT} and V_{CTAT}, respectively. The difference of V_{PTAT} and V_{CTAT} is used to measure temperature in the back-end-read-out circuits shared across front ends. The equation for V_{PTAT} can be derived by simply equating the current flowing through M1 and M2. Since both the transistors are in the deep subthreshold region (M1 due to high V_{TH} and M2 due to shorted gate and source), subthreshold current equations can be applied on both the transistors and then equated and solved for V_{PTAT} as shown below.

$$I_{D1} = \mu_1 C'_{ox1} \frac{W_1}{L_1}(n_1 - 1)\phi_t^2 \cdot \exp\left(\frac{V_{PTAT} - V_{TH1}}{n_1 \phi_t}\right)$$
$$I_{D2} = \mu_2 C'_{ox2} \frac{W_2}{L_2}(n_2 - 1)\phi_t^2 \cdot \exp\left(-\frac{V_{TH2}}{n_2 \phi_t}\right). \quad (2)$$

$$V_{PTAT} = \underbrace{n_2 \ln\left(\frac{\mu_1}{\mu_2} \cdot \frac{C'_{ox1}}{C'_{ox2}} \cdot \frac{W_1 L_2}{W_2 L_1} \cdot \frac{n_1 - 1}{n_2 - 1}\right) \frac{k}{q} . T}_{\text{temperature slope}} + \underbrace{V_{TH2} - \frac{n_2}{n_1} V_{TH1}}_{\text{offset}} \quad (3)$$

The fact to be noted here is that the offset term in equation (3) itself contains V_{TH} term which will cause nonlinearity even in the differential readout of V_{PTAT} and V_{CTAT}. The measurement results of 400 μm^2 front

end show a worst-case error of 5.4 °C across 64 instances. The circuit consumes 0.85 µW of power for the front end design and a total of 0.36 mW power including the readout circuitry. This sensor technique requires additional ADC to convert the analog voltage to equivalent digital temperature data. This design is also affected by local process variations in differential read-out mechanism.

One of the major concerns with these analog designs is their large area and current to be accompanied by digital circuits working on low supply voltages. To overcome this problem, Ring Oscillator (RO) based designs with transistors working in the deep subthreshold region were proposed. The additional advantage of this approach is that along with temperature these can be tuned to sense process variations as well. In [4] a Ring oscillator based temperature sensor is proposed which doesn't require a reference clock and has reduced supply sensitivity. This sensor harvests the V_{TH} dependency of VCO frequency and converts temperature into the frequency and then into digital data for further processing. The sensor works with the supply voltage varying from 0.85 V to 1.05 V with a supply sensitivity of 0.034 °C/mV. The basic working architecture is shown in Fig 1.4 below.

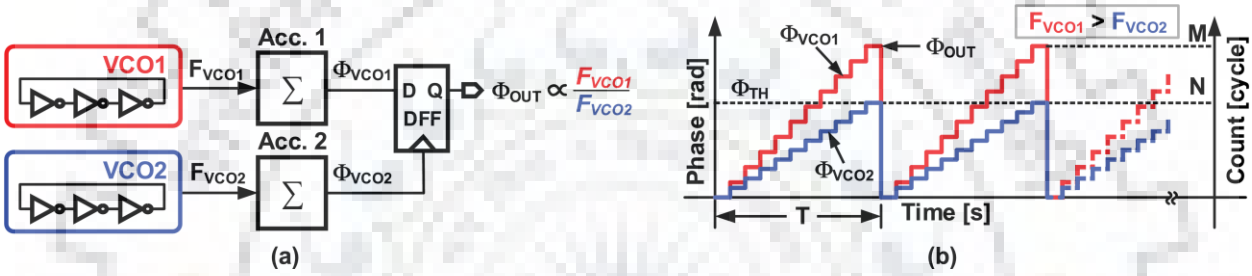


Fig. 1.4 (a) Basic architecture of the proposed temperature sensor. (b) The phase versus time plot showing division concept in the proposed architecture. [4]

Two RO chains with different V_{TH} sensitivities are used such that the frequencies of both RO chains have different CTAT slopes with temperature. Their frequencies are accumulated and divided such that the ratio of F_{VCO1} and F_{VCO2} is PTAT. The ratio F_{VCO1}/F_{VCO2} is independent of µ and only dependent on V_{TH}, although F_{VCO1} and F_{VCO2} individually are dependent on both µ and V_{TH}. The frequency ratio is given by equation (4).

$$\frac{F_{VCO1}}{F_{VCO2}} \alpha \frac{(V_{DD} - V_{TH1})\, C_{L2}}{(V_{DD} - V_{TH2})\, C_{L1}} \tag{4}$$

Reverse short channel effect is leveraged to create the required threshold voltage difference between the transistors used in the delay cells of VCO1 and VCO2. Transistors in VCO2 have longer channel lengths as compared to the transistors in VCO1. The advantages of this approach over other previous approaches are that it is a complete digital technique ready to be operated even with logic supply voltage. These RO based sensors don't require a clock or bandgap reference. Also, the sensor can be made insensitive to supply variations by trimming junction capacitance of inverter. The layout of this design occupied an active die area of 0.004 mm^2 in 65 nm CMOS technology. The peak to peak nonlinearity with and without polynomial correction is ± 0.9 °C and 2.3 °C, respectively, over a temperature range from 0 °C to 100 °C.

Chapter 2: Proposed Approach

2.1 Temperature Sensor Concept

In the proposed approach RO chain works as a temperature sensing element. Change in the temperature near this sensing element changes the V_{TH} and μ of the NMOS and PMOS embedded in each delay stage of RO chain. Changed V_{TH} and μ changes the current which charges and discharges the load which in turn changes the oscillation frequency of RO. Mathematically the frequency of an RO (F_{RO}), to the first order approximations, can be easily expressed by the simple equation as

$$F_{RO} \propto \frac{4}{3} \frac{\mu C_{OX} \frac{W}{L} (V_{DD} - V_{TH})^2}{V_{DD}(1 - \frac{5}{6}\lambda V_{DD}) C_L} \quad (1)$$

Here μ is mobility of electrons/holes, W and L are the width and length of the MOS transistors, C_{OX} is the gate-oxide capacitance, V_{DD} is the supply voltage, V_{TH} is the average threshold voltage of NMOS and PMOS transistors used in the single delay stage, λ is the channel length modulation parameter and C_L is the load capacitance of single delay stage.

Mobility and threshold voltage have CTAT and PTAT temperature dependency respectively, which can be expressed as

$$\mu \propto \mu_0 \left(\frac{Temp}{T_0}\right)^{-p} \quad (2)$$

$$V_{TH} = V_{TH0} - k(Temp - T_0)$$

Where μ_0 is the mobility at room temperature, p is a fitting parameter generally in the range of 1.2 to 2.0, and k is approximately in the range of 1-3 mV/ °C.

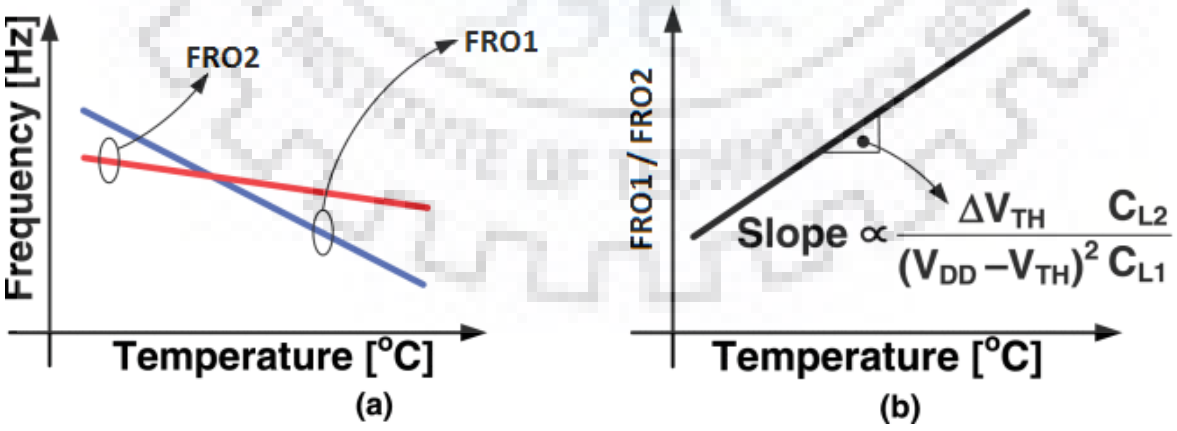


Fig. 2.1 (a) Frequency of RO1(F_{RO1}) and RO2(F_{RO2}) versus temperature. (b) The ratio of the frequency of RO1 over frequency of RO2 (F_{RO1}/F_{RO2}) versus temperature.

A single RO is sufficient to measure temperature changes through its frequency but this CTAT curve is not very linear because F_{RO} is simultaneously dependent on μ as well as V_{TH}. Thus one of the dependencies needs to be removed from the F_{RO} equation, which can be easily done by taking the ratio of two RO frequencies. From the equations of μ and V_{TH}, it can be noted that μ has a CTAT relation with temperature while V_{TH} has a PTAT relation. So the temperature sensitivity of an RO can be modified by either changing the mobility or threshold voltage. Since we cannot control mobility directly, V_{TH} is used to create temperature sensitivity difference.

The proposed temperature sensor architecture has a single RO chain having 17 stages of configurable delay stages. These delay stages are designed such that they can show two different kinds of delays depending upon the status of the select signal. So two different kinds of frequencies i.e. F_{RO1} and F_{RO2} can be obtained from single RO chain multiplexed in time with select signal. The difference in sensitivity is created using two different inverters having different threshold voltages. Inverter stages which are producing F_{RO1} have high threshold voltage compared to inverter stages which produce F_{RO2}. As a result, RO2 frequency is more dependent on the effect of mobility variation due to temperature compared to RO1 frequency. Consequently, the frequency versus temperature plot of RO2 has a steeper slope as compared to RO1 in Fig. 2.1(a).

$$\frac{F_{RO1}}{F_{RO2}} \alpha \frac{(V_{DD} - V_{TH1})\, C_{L2}}{(V_{DD} - V_{TH2})\, C_{L1}} \tag{3}$$

The ratio of frequencies of RO1 and RO2 [F_{RO1}/F_{RO2}] shows the desired PTAT characteristic as in Fig. 2.1(b) and depicted from equation (3). This ratio can be digitized with the help of digital logic to obtain the digital output carrying the temperature information.

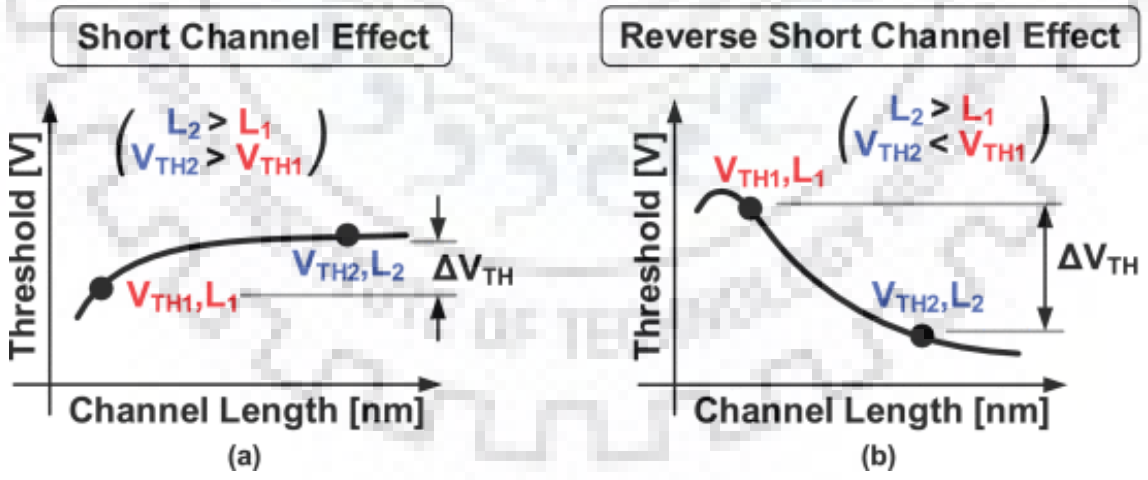


Fig. 2.2 (a) Effects of channel length on threshold voltage due to short channel effects. (b) Effects of channel length on threshold voltage due to reverse short channel effects.

As the channel length of a transistor decreases, short channel effects reduces its threshold voltage (refer Fig. 2.1). On the other hand, in case of reverse short channel effect threshold voltage increases as the channel length decreases (refer Fig. 2.2). Since in our technology node i.e. 65 nm RSCE is present therefore we use it to create the threshold voltage difference. RSCE is dominant in 65 nm technology node and is therefore leveraged to create threshold voltage difference between two transistors by increasing the channel length of one transistor compared to other. In advanced technology nodes in which RSCE is no longer present, two different types of transistors should be used to create the required threshold voltage difference.

In this work, we have used longer channel length transistors in RO2 as 600 nm and 200 nm channel length in RO1 to induce a threshold voltage difference of about 27 mV in NMOS transistors and 19 mV in PMOS transistors.

2.2 Supply Voltage Variations Sensitivity

One of the key features of the on-chip temperature sensor is that it should be insensitive to supply voltage variations. Since in a single chip there are numerous sensors placed so in order to reduce the overhead associated with the routing of separate dedicated power supply, a sensor should be able to operate from the logic supply voltage. However, because of the presence of DVS in modern processors and high switching there tend to be low average voltage and fluctuating supply respectively. Therefore, a sensor should be independent of supply voltage variations.

As stated earlier the frequency of a ring oscillator (RO) is given by equation (4) below, using which ratio

$$F_{RO} \propto \frac{4}{3}\frac{\mu C_{ox} W/L (V_{DD} - V_{TH})^2}{V_{DD}\left(1 - \frac{5}{6}\lambda V_{DD}\right) C_L} \tag{4}$$

of the frequencies of two RO chains can be expressed by equation (5) –

$$\frac{F_{RO1}}{F_{RO2}} \propto \frac{(V_{DD} - V_{TH1})^{\alpha} C_{L2}}{(V_{DD} - V_{TH2})^{\alpha} C_{L1}} \tag{5}$$

Here V_{TH1} and V_{TH2} are threshold voltages of transistors in RO1 and RO2, C_{L1} and C_{L2} are load capacitances of the delay stages in RO1 and RO2, and α is from the α-power law model [1]. It is quite clear from equation (5) the frequency ratio F_{RO1}/F_{RO2} is dependent on the supply voltage.

The load capacitance C_L of an inverting stage (delay cell) consists of gate-to-source capacitance (C_{GS}) of next stage, gate-to-drain capacitance (C_{GD}) of next stage, wire capacitance (C_W) and drain-to-bulk capacitance (C_{DB}) of same stage [6]. Here C_{DB} is the reverse biased PN junction which reduces when the reverse bias voltage across the PN junction increases as shown in Fig. 2.3. Since body is tied to ground and V_{DD} in NMOS and PMOS respectively, so effectively it is the drain voltage which is going to affect this capacitance.

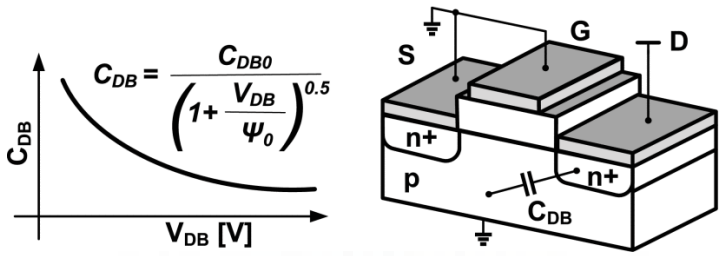


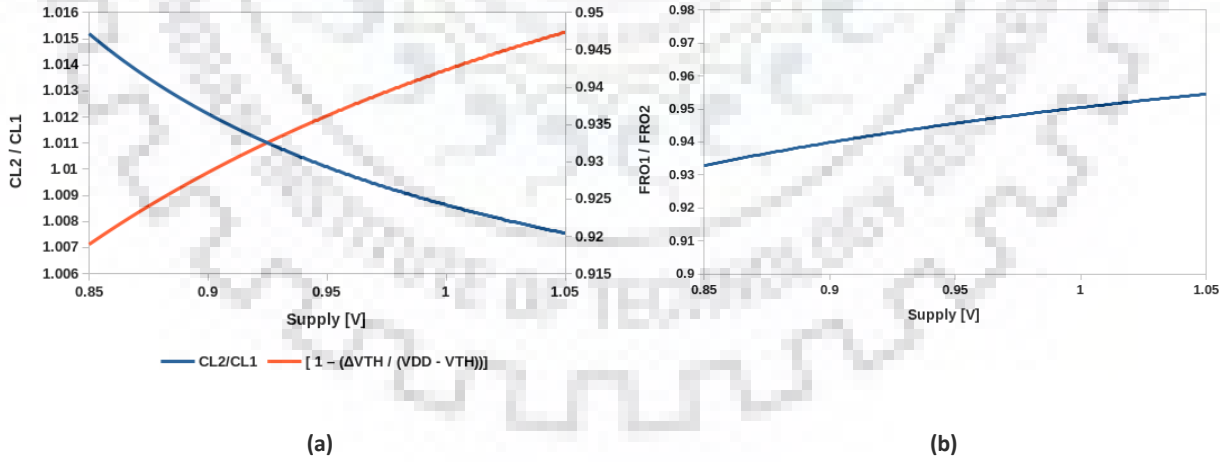
Fig. 2.3. Drain to body junction capacitance (C_{DB}) versus reverse bias voltage.

After simplification of equation (5) a new equation (6) is obtained which has two terms. The First term which is inside the square brackets, shows a positive dependency on supply voltage i.e. F_{RO1}/F_{RO2} increases as the supply voltage increases as shown in Fig. 2.4(a).

$$\frac{F_{RO1}}{F_{RO2}} \propto \left[1 - \frac{\Delta V_{TH}}{V_{DD} - V_{TH}}\right]^{\alpha} \frac{C_{L2}}{C_{L1}} \tag{6}$$

$$\frac{C_{L2}}{C_{L1}} \propto \left[1 + \frac{\Delta \Psi}{V_{DD} + \Psi}\right]^{0.5} \tag{7}$$

The second term is ratio C_{L2}/C_{L1} which shows a negative sensitivity to supply voltage given by equation (7) and illustrated by Fig. 2.4(a), considering that C_{DB} is dominant in load capacitance C_L in a delay stage. So if load capacitance C_L is changed by changing C_{DB} such that it cancels out the effect of the first term then frequency ratio F_{RO1}/F_{RO2} can be obtained as independent of supply voltage as shown in Fig. 2.4(b).



(a) (b)

Fig. 2.4 (a) Load capacitance dependence on supply. (b) Frequency ratio with supply after matching load capacitance with first term in equation (6).

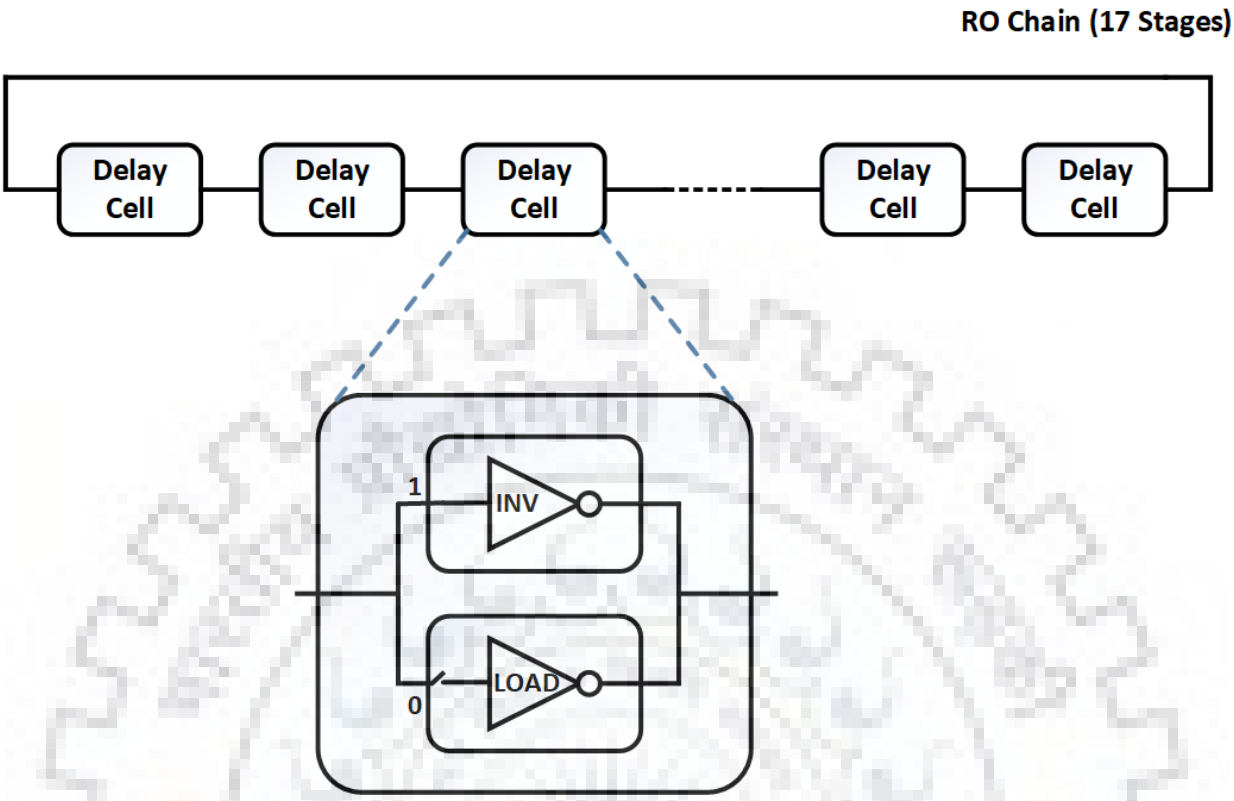


Fig. 2.5. Block diagram representation of proposed temperature sensor.

In the proposed sensor architecture, a single delay cell is designed such that the C_{DB} dominates among all the capacitances contributing in the total load capacitance. The size of the transistors in load stage is chosen such that the effective C_{DB} and hence C_{L2}/C_{L1} due to cancels out the supply sensitivity of the first term in equation (3). Consequently, the ratio F_{RO1}/F_{RO2} becomes more tolerant to variations in the supply voltage.

The proposed circuit is shown in Fig. 2.5, in this circuit, when RO1 behaves as an inverter, RO2 behaves as a dummy load in each stage of RO chain, such that its effective load capacitance is increased. Since we have to increase the ratio C_{L2}/C_{L1} to cancel the effect of the first term of equation (3), that is why the dummy load is added in F_{RO2} chain only. The drain-to-body capacitance (C_{DB}) is directly proportional to width (W) of the transistor thus increasing the width (W) of dummy NMOS and PMOS increases C_{L2}/C_{L1} factor and supply sensitivity seems to be decreasing. When RO2 acts as inverter RO1 acts as inverter but since its size is very less compared to RO2, it doesn't have any significant load characteristics for RO2.

2.3 Sensor Architecture

The proposed temperature sensor uses a single RO chain instead of two separate ROs, to create two different frequencies. This RO consists of 17 stages of reconfigurable delay cells such that they produce one frequency, F_{RO2} when the select line is low and second frequency (F_{RO1}) when the select line is high. 2x1 MUXs designed using pass transistors are used to let the structure switch between inverter or load. The proposed sensor converts temperature information to the frequency and the digitized frequency ratio

provides temperature information in digital bits. The oscillation frequency of CMOS ring oscillators (RO) is dependent upon temperature through mobility (μ) and threshold voltage (V_{TH}) variations of transistors. A division operation will be carried out on F_{RO1} and F_{RO2} in a time multiplexed fashion.

As shown in Fig. 2.6 MUX1 & MUX2 have a common select line, so when select line S is given logic zero, transistors M1 and M2 receive a common signal i.e. Input and this structure behaves as simple inverter. At the same time, MUX3 & MUX4 pass V_{DD} & GND to transistors M3 & M4 respectively because of which they are turned off and behave as the load on M1 and M2.

When S is switched to logic one, M1 and M2 receive V_{DD} & GND respectively and behave as the load to M3 & M4 which in turn are now working as normal inverter since MUX3 & MUX4 pass input to both the transistors.

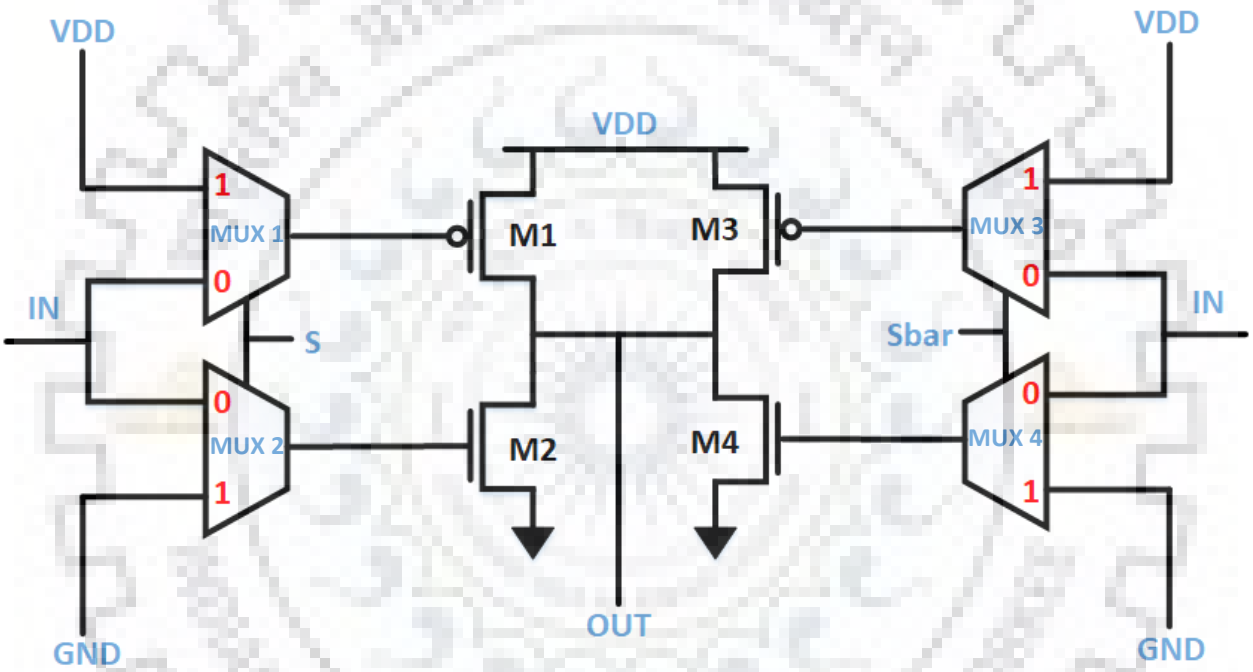


Fig. 2.6. Circuit diagram of a single delay cell of the proposed approach.

Temperature sensitivities have been modified by changing the length of transistors M1 & M2 from M3 & M4. M1 and M2 have aspect ratio, W/L = 400nm/200nm and 200nm/200nm respectively while M3 & M4 have aspect ratio W/L = 9μm/0.06μm & W/L = 9μm/0.06μm respectively. The number of the stages are same i.e. 17 for both the cases of select line. F_{RO1} has high value and steep slope because a large amount of current flowing through M3 & M4 charges and discharges a small load of M1 and & M2 compared to F_{RO2} where a small amount of current flowing through weak M1 & M2 charges and discharges a strong load of M3 & M4. Thus F_{RO1} is quite higher than F_{RO2} and hence temperature sensitivity of design is 0.0048/°C, which is much better compared to supply sensitivity of [4], i.e. 0.0017/°C.

The advantage of this technique over the previous [4] is based on the basic architecture only. Since there is only one RO chain in this model there are no inter-RO process and mismatch variations in mobility (μ) and threshold voltage. Only with-in RO variations are present among different inverters in the same RO

chain. Also, we require load stage to balance supply sensitivity, so we are using it to create frequency as well. In this model frequency ratio does not depend on the number of stages, it remains constant for any number of stages while in [4] frequency ratio depends on the number of stages in chain 1 and 2. Due to spatial correlation, intra-RO process and mismatch variations are also minimized in this approach. Although a single delay cell here occupies more area compared to that of [4], but in terms of application and scalability this approach is more area efficient. For example, to reduce the frequency by half then 16 stages should be added in the existing design, while in [4] you need to increase 34 stages in chain 1 and 66 stages in chain 2, i.e. 100 stages in total to keep frequency ratio constant. Thus overall this approach solves quite a few architectural level problem of design in [4].

2.4 Summary

In this chapter a new model is proposed for sensing temperature. Although the concept of this design is same to what is mentioned in [4], but the architecture of our design is a bit different from [4]. In our model a single delay cell is used to generate two different frequencies on same output pin. A select line is used to switch between both the frequencies. The design is also made insensitive to supply variations by balancing the load capacitance with V_{DD} dependence of frequency ratio. The design used multiplexers for switching between frequencies. These multiplexers are designed using the pass transistors logic and minimum dimensions so as not to load the input much.

Chapter: 3 Simulation Results

3.1 Post Layout Simulations

Layout for the proposed temperature sensor is made in a 65 nm CMOS technology node and is simulated post layout. The model operates with temperature varying from 0-100 °C and a supply voltage range of 0.85 V to 1.05 V (200 mV). All the simulations are done at a supply voltage of 1.0 V and temperature of 30 °C unless otherwise stated. The sensor occupies a total area of 0.00053 mm^2, as shown in the layout in Fig. 3.1(a). The transient response of the sensor is shown in Fig. 3.1(b). When transiting from high frequency to low frequency, this technique takes some time to settle down the output to the low frequency.

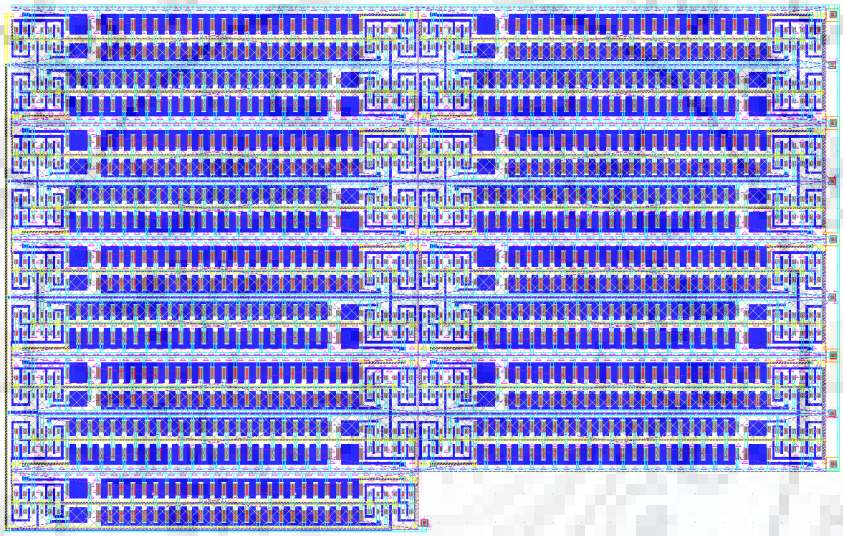


Fig. 3.1. (a) Layout of proposed design in 65 nm CMOS technology.

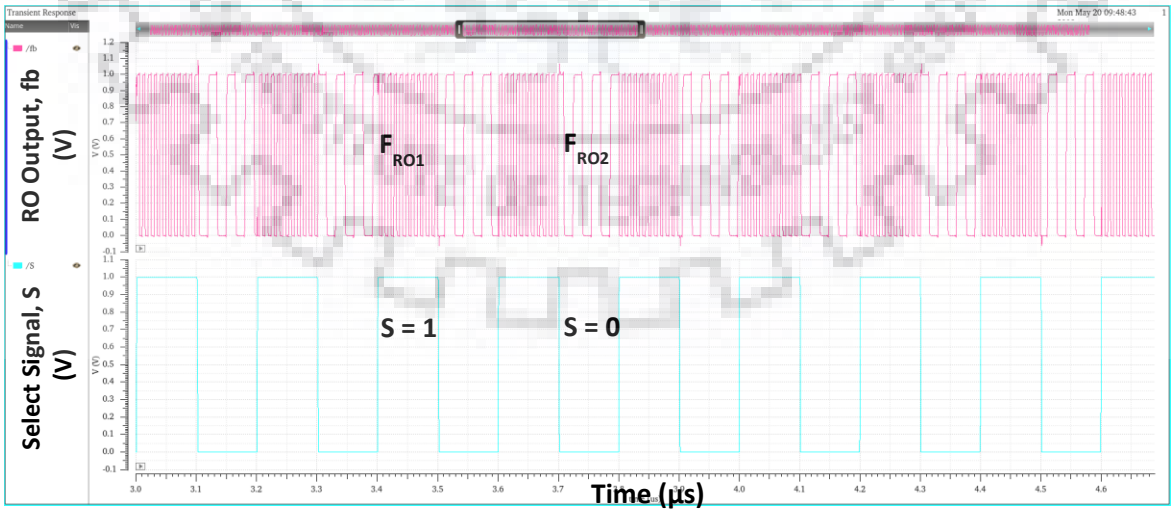


Fig. 3.1. (b) Transient response of the proposed design.

The temperature sensitivity of F_{RO1} and F_{RO2} is plotted in Fig. 3.2(a) and (b) respectively. F_{RO1} varies from 86 MHz to 79.1 MHz having a slope of −69 KHz/°C for a temperature variation of 0 °C to 100 °C. F_{RO2} varies from 24.66 MHz to 19.93 MHz with a slope of approximately −47.3 KHz/°C for a temperature change from 0 °C to 100 °C. As described earlier, the negative slope of RO1 is steeper compared to the negative slope of RO2 because of the aspect ratio and hence strength of transistors in the inverter of RO1 are much higher compared to that of RO1. The ratio of frequencies of two ROs i.e. F_{RO1}/F_{RO2} has the PTAT characteristics, as shown in Fig. 3.2(c).

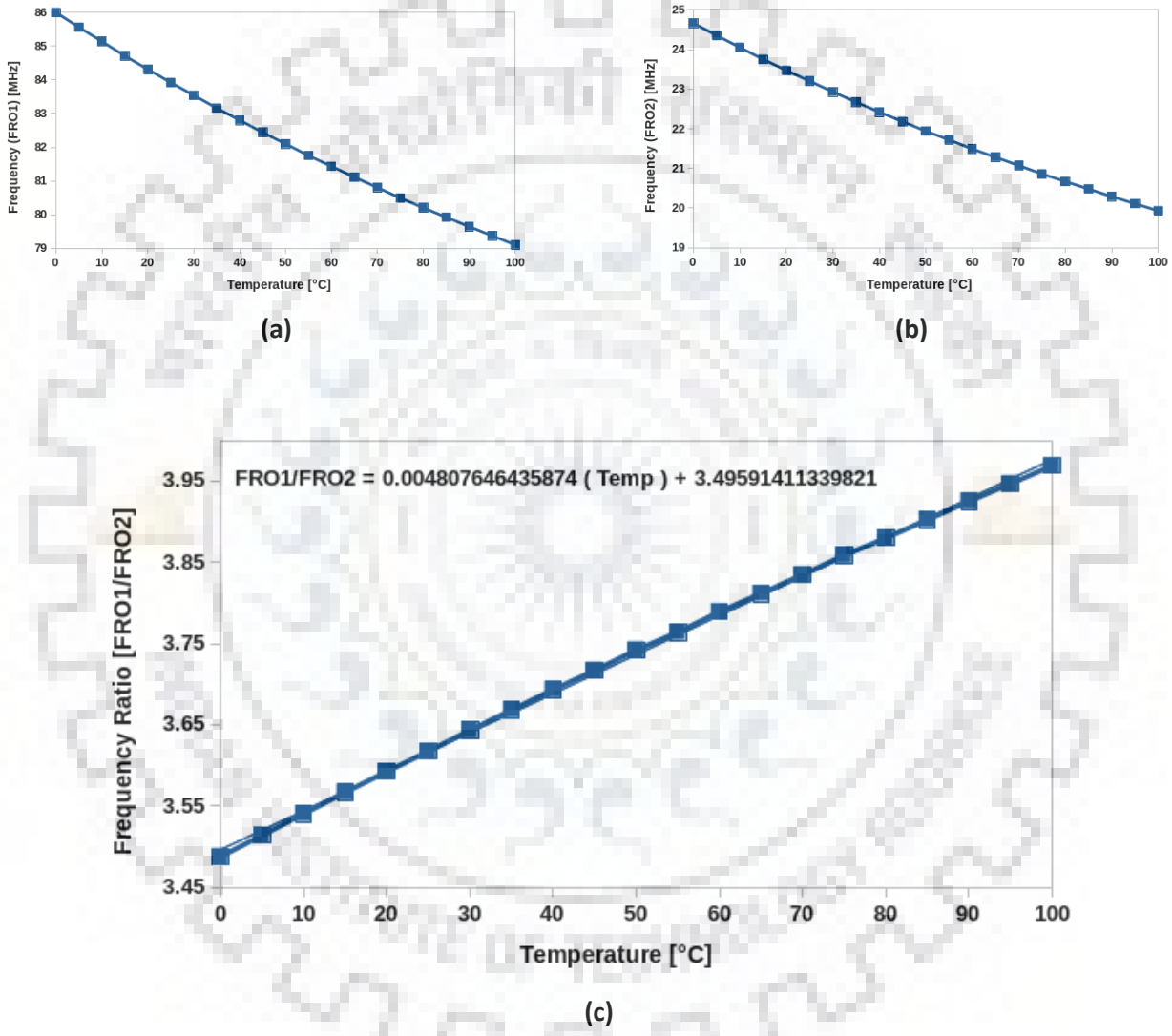


Fig. 3.2 (a) Simulated RO1 frequency (F_{RO1}) versus temperature. (b) Simulated RO2 frequency (F_{RO2}) versus temperature. (c) Simulated frequency ratio of RO1 frequency over RO2 frequency (F_{RO1}/F_{RO2}) versus temperature.

The errors in temperature linearity of the proposed sensor with TPC is shown in Fig. 3.3. With two-point calibration, at 30 °C, the peak-to-peak nonlinearity is 2.87 °C, without polynomial correction over a temperature range of 0 °C to 100 °C.

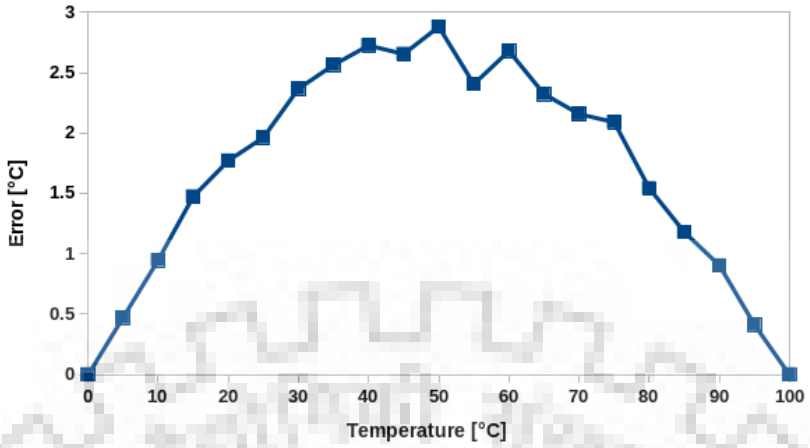
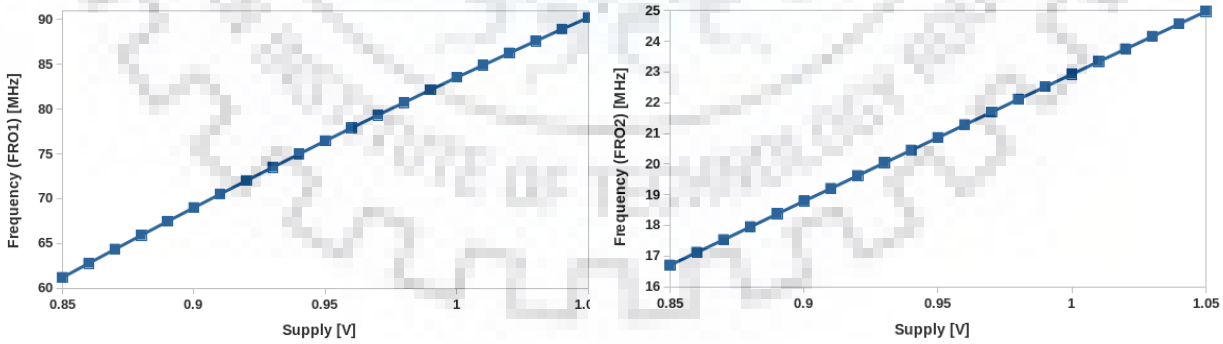


Fig. 3.3 Non-linearity in frequency ratio versus temperature expressed in terms of error in temperature after two-point calibration at 0 °C and 100 °C.

The supply sensitivity of F_{RO1} and F_{RO2} is plotted in Fig. 3.4(a) and (b) respectively. RO1 frequency shows a variation from 61.21 MHz to 90.22 MHz having a slope of 145.05 KHz/mV for a supply voltage range of 0.85 V to 1.05 V (200 mV). RO2 frequency shows a variation from 16.7 MHz to 24.9 MHz having a slope of 41 KHz/mV for a supply voltage range of 0.85 V to 1.05 V (200 mV). The sensitivity of load capacitance (C_{DB}) of RO2 to supply voltage helps to match the supply sensitivity of the two oscillators. The error in frequency ration of sensor due to supply voltage variation was simulated at 30 °C, and the results are shown in Fig. 3.5(b). For these measurements, the supply voltage was varied by 200 mV, ranging from 0.85 V to 1.05 V. At 30 °C, the measured peak-to-peak variation is 6.55 °C for 200 mV of DC supply voltage variation, which is equivalent to a supply sensitivity of 0.032 °C/mV. In the current implementation, the junction capacitance (C_{DB}) is fixed, which resulted in limited supply noise cancellation. However, if this capacitance is balanced depending on the operating supply voltage, supply sensitivity can further be reduced over a range of temperature and across process corner.



(a) **(b)**

Fig. 3.4 (a) Simulated RO1 frequency (FRO1) versus supply. (b) Simulated RO1 frequency (FRO1) versus supply.

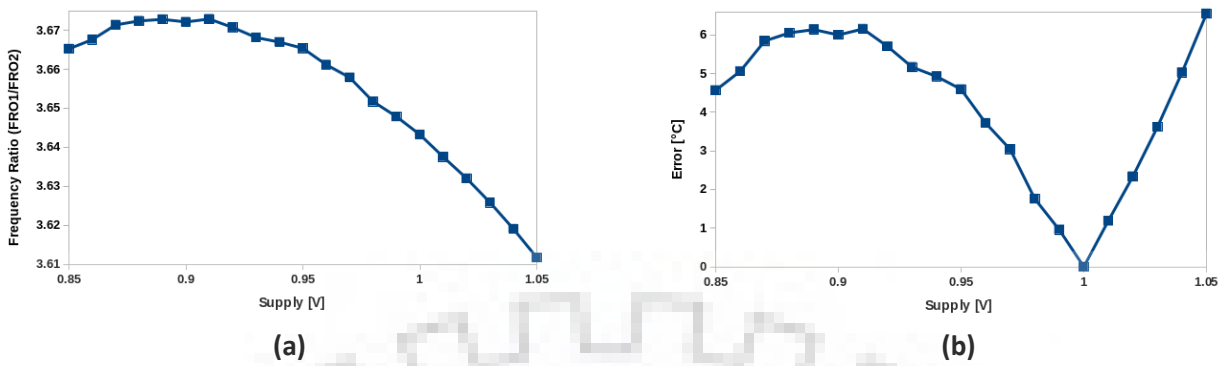


Fig. 3.5 (a) Simulated frequency ratio of RO1 frequency and RO2 frequency (F_{RO1}/F_{RO2}) versus supply. (b) Simulated temperature sensor error versus supply voltage at 30 °C.

In order to check for process and mismatch variations 1000 Monte Carlo simulations were performed on the design with only mobility (μ) variations and keeping threshold voltage (V_{TH}) and other parameter variations constant. Simulation is done around the typical corner (TT) at 30 °C for 100 ns transient time to find F_{RO1}, F_{RO2}, and F_{RO1}/F_{RO2}. Histogram resulted from Monte Carlo simulations for F_{RO1}/F_{RO2} is shown in Fig. 3.6.

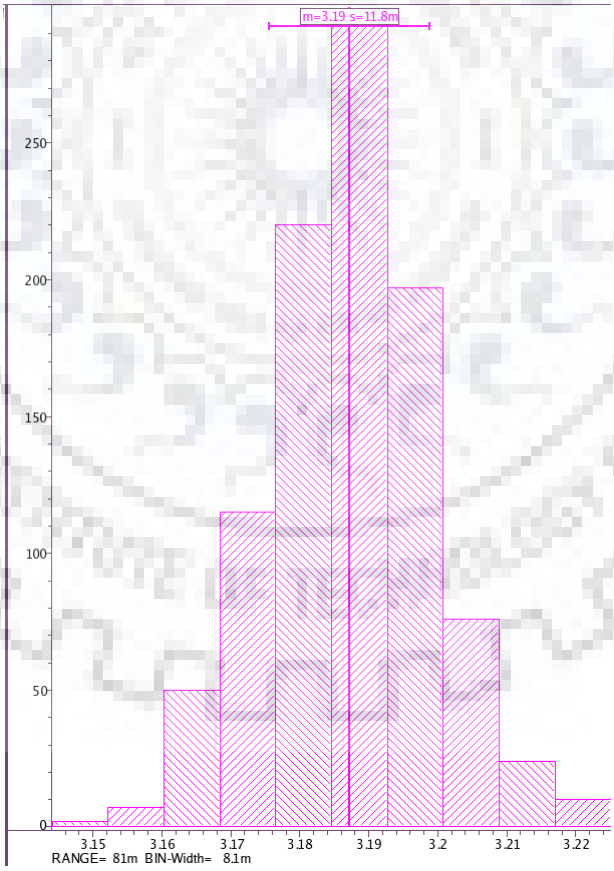


Fig. 3.6 Frequency ratio sensitivity with 1000 Monte Carlo Simulations at TT corner at 30°C with mobility variations.

3.2 Comparison Table

	This Work	JSSC'16 [4]
Technology	65 nm	65 nm
Type	MOSFET	MOSFET
Area [mm^2]	0.00053	0.00014
Supply [V]	0.85-1.05	0.85-1.05
Power	180 µW @ 1V	154 µW @ 1V
External Clock Reference	YES	NO
Supply Regulator	NO	NO
Temperature Range [°C]	0-100°C	0-100°C
Calibration	2-point	2-point
Temperature Sensitivity [1/°C]	0.0048 / °C @ 1V	0.0017 / °C @ 1V
Temperature Inaccuracy (w/o correction)	2.87 °C @ 1V	6.47 °C @ 1V
Supply Sensitivity [°C/mV]	0.032 °C/mV @ 30°C	0.055 °C/mV @ 30°C
Supply Inaccuracy	6.55 °C	11.07 °C
Temperature Error (due to mismatch)	2.6 °C	4.43 °C

Table 3.1 Performance comparison of the proposed temperature sensor with reference paper [4].

3.3 Summary

Layout for the proposed design was made in 65 nm CMOS technology and post-layout simulations were done to extract temperature and supply sensitivities as shown. The layout occupied an active area of 0.00053 mm^2. The temperature and supply sensitivities of our design are better than that of design in [4], which is also made in 65 nm CMOS technology node. Linearity error in temperature is calculated using two-point calibration and Monte Carlo simulation were run at TT corner to study the effects on frequency ratio due with-in die process variations in mobility. The power consumption in case of our design is nearly equal to the power consumption in [4].

Chapter 4: Dynamic Voltage Scaling Through Die Temperature Feedback

Dynamic voltage scaling is a technique in which the supply voltage of a system is dynamically varied depending upon some feedback from a system such that system temperature and power consumptions are within a certain limit. In this design, the temperature is used as the feedback to scale supply voltage of FPGA peripherals, so as to limit the die temperature from going beyond a limit of 40 °C.

4.1 System Level Implementation

Dynamic Voltage Scaling is implemented on a VC707 evaluation board which has the XC7VX485T-2 FPGA on it. The block diagram of the circuit that is used to dynamically scale the voltage of the FPGA is shown in Fig. 4.1. Different parts of the block diagram are PicoBlaze, pre-programmed ROM, Interrupt controller, XADC, adders, multiplexer, demultiplexer, deciding logic, power regulator and PMBus controller. The supply voltage is regulated by the power regulator depending on the commands coming from the PicoBlaze through the PMBus. These commands are generated based on the die temperature, calculated using ADC and interrupt controller. This forms the closed loop configuration in which the voltage is controlled by the feedback from the die temperature.

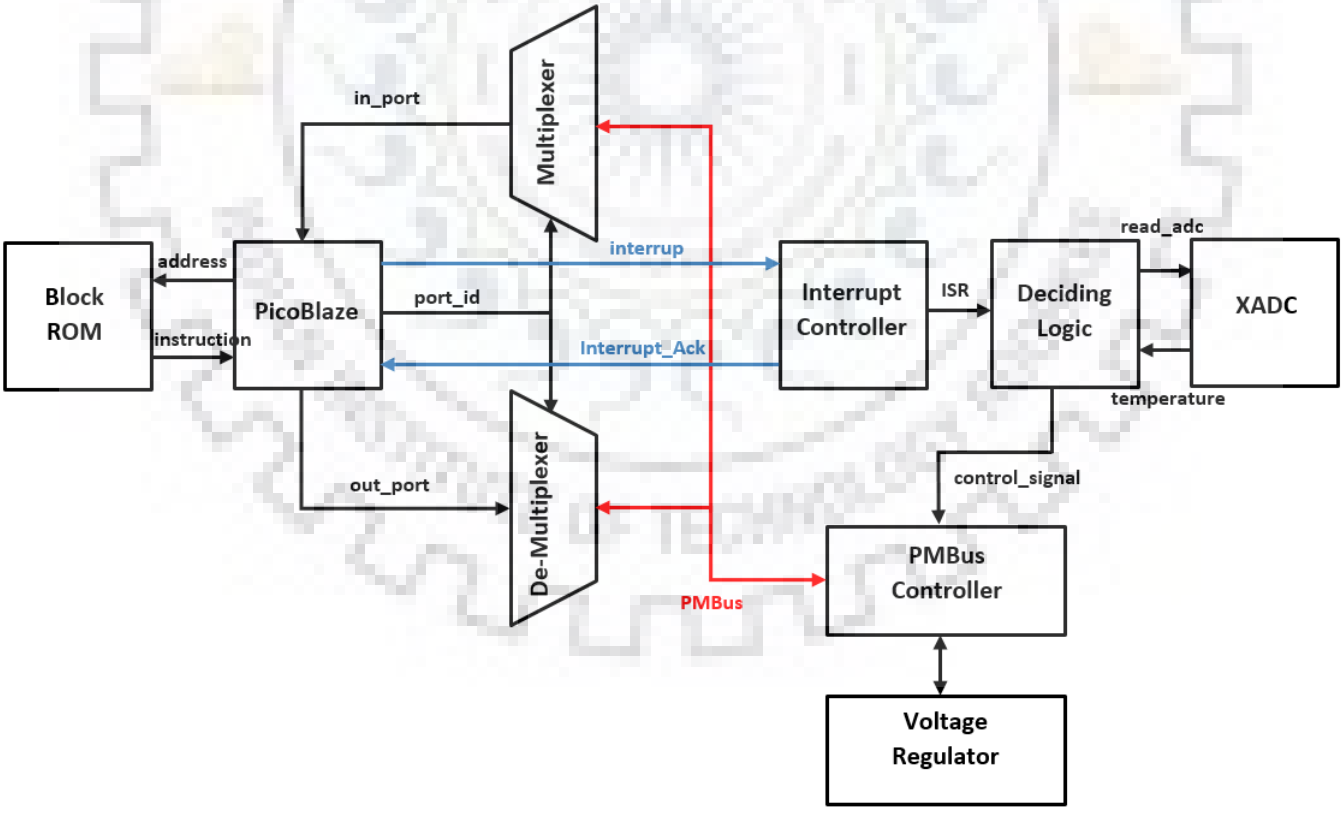


Fig. 4.1 System level implementation of Dynamic Voltage Scaling.

4.2 PicoBlaze

PicoBlaze is an 8-bit RISC microcontroller implemented as a soft macro for use in FPGA's. It can be included one or more times in any of the Virtex-6, Spartan-6, 7-series or Zynq designs. It uses a block memory and 26 logic slices. The main advantage of using a PicoBlaze in any design is that hardware is parallel and processors are sequential. Hence implementing a processor is very much useful when a large number of states are to be achieved in a particular sequential order. This is also useful to time-share hardware resources when there are several other slower tasks to be performed.

The PicoBlaze processor can execute any program of up to a maximum of 4K instructions. Every instruction in PicoBlaze takes 2 clock cycles to execute and every instruction is 18-bit long. This makes the flow of the program very predictable and the time to calculate the completion of a routine can be easily estimated. There are a total of 68 opcodes which are used in PicoBlaze. The maximum clock frequency is device dependent. A maximum clock frequency of 238 MHz is achieved in the Kintex-7(-3 speed grade) device. The block diagram of PicoBlaze is shown in Fig. 4.2.

PicoBlaze has 2 banks of 16 general purpose registers. These can be used to manipulate the flow of 8-bit data. The ALU (Arithmetic and Logical Unit) provides all the basic comprehensive set of instructions for performing AND, OR, ADD, SUB, left shift, right shift, rotate, TEST, COMPARE. The instructions JUMP, CALL, RETURN control the flow of the execution.

The other advantages of using a microcontroller within the FPGA fabric over other FPGA's like Zynq which have a full-fledged microprocessor is that it avoids the complex ways of communicating between the processor and the FPGA fabric. The PicoBlaze is instantiated as a module and hence the communication is simplified by using the output pins as a wire.

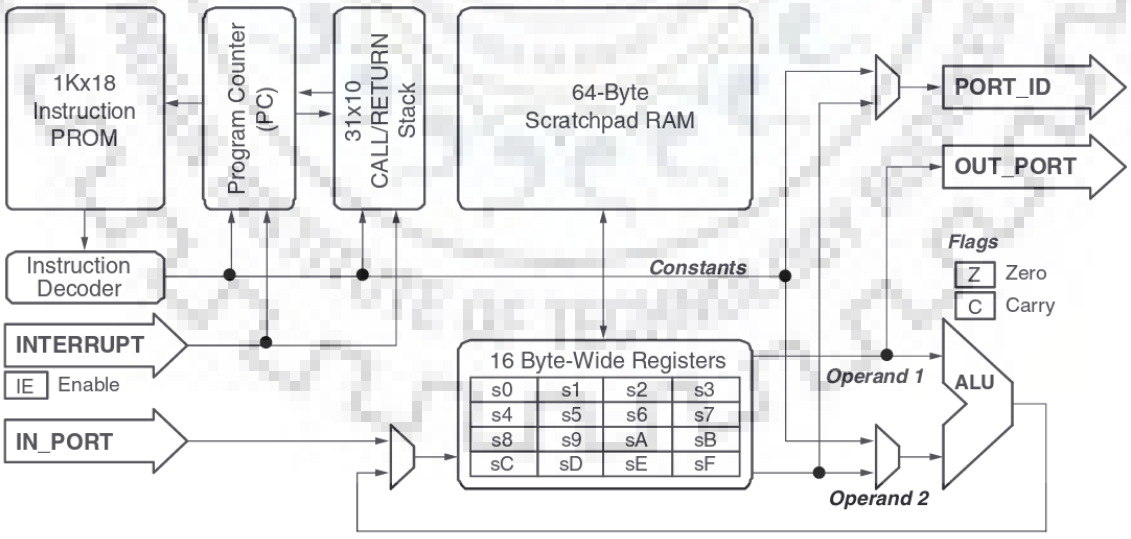


Fig. 4.2 Block diagram of PicoBlaze.

The sequential program to be run on PicoBlaze is written in assembly language and stored in a file with ". psm" extension. This program is compiled using an assembler to generate the Verilog or VHDL file which contains the instructions to be stored in ROM (Read Only Memory) and can be used directly in the top-level implementation of the design. This Verilog or VHDL module is interfaced with the PicoBlaze module and the instructions are read from this ROM during the execution of the design.

The PicoBlaze microcontroller using the OUTPUT instruction can output 8-bit values to up to 256 general purpose output ports. This is implemented using port_id to which the output port is written and the data is written to the out_port. The write_strobe signal is set high when writing data to an output port. Similarly, the input is taken from the INPUT instruction by writing the input port to port_id and taking the input in the in_port.

The PicoBlaze microcontroller also has an optional interrupt input which allows for handling external events asynchronously. The Interrupt Service Routine (ISR) is called within five clock cycles. The interrupt_ack pin is set high to acknowledge that the interrupt is received properly and being serviced.

4.3 Power Rails

Dynamic Voltage Scaling is done on a VC707 evaluation board which has a Virtex-7 XC7VX485T-2 FPGA on it. There are several power rails present on this board that supply power to various parts such as the FPGA core, I/O, etc. The voltage to the FPGA core is supplied through the power rail named VCCINT. Changing the value of VCCINT changes the voltage supplied to the FPGA which in turn changes the temperature of the die of FPGA. The main supply to the VC707 board is a 12V external supply. The supply to the power rail VCCINT is adjusted using a switching regulator PTD08A020W present on the board. This switching regulator is controlled by UCD9248 PMBus controller from Texas Instruments present on the board. Table 4.1 shows all the power rails that can be controlled using three UCD9248 PMBus controllers and the addresses of the PMBus at which these controllers can be accessed.

	Rail number	Rail name	Nominal voltage(V)
UCD9248 Power Controller at Address 52	Rail #1	VCCINT	1
	Rail #2	VCCAUX	1.8
	Rail #3	VCC3V3	3.3
	Rail #4	VADJ	1.8
UCD9248 Power Controller at Address 53	Rail #1	VCC2V5	2.5
	Rail #2	VCC1V5	1.5
	Rail #3	MGTAVCC	1
	Rail #4	MGTAVTT	1.2

UCD9248 Power Controller at Address 54	Rail #1	VCCAUX_IO	2
	Rail #2	VCC_BRAM	1
	Rail #3	MGTVCCAUX	1.8
	Rail #4	VCC1V8	1.8

Table 4.1 Different Power rails connected to PMBus Controllers.

4.4 PMBus

I2C bus (Inter-Integrated Circuit) is a packet switched, multi-master, multi-slave serial bus. The communication in I2C occurs through two wires namely SCL (Serial Clock) and SDA (Serial Data) which are pulled up to V_{CC} through a resistor. The current bus master generates the clock signal and the data is sampled by the slave devices at the positive edge of this clock signal. The message format consists of two frames, address frame, and data frame. In the address frame, the master specifies the slave to which the message is being sent and the data frame consists of one or more 8-bit data packets sent either from the slave to master or vice versa. SMBus (System Management Bus) is a subset of the I2C bus. I2C protocol supports arbitrary message structures while the SMBus restricts the number of message structures to nine.

PMBus refers to the Power Management Bus. It is a variant of SMBus (System Management Bus) which helps in managing the power supplies. It defines a number of domain-specific commands required by power control and management components which have to be implemented along with the communication protocol. The PMBus slave must be able to start up in a controlled manner without any interaction from the serial bus. This ensures that the power is supplied properly with the default values even when the master is unable to communicate with the slave.

PMBus has predefined set of commands in comparison with SMBus in which the user defines the set of commands that are to be operated. There are a total of 255 sets of commands given in the PMBus specification out of which 208 commands are standardized for all manufacturers and products, whereas the rest of the commands are set by the manufacturer of the product using the PMBus. There are a variety of PMBus commands which are used to monitor and/or manipulate the device attributes such as voltage, current levels, temperature, etc. For example, the PAGE command is used to select a power rail on which subsequently we can operate other commands, like reading or setting the values of voltage, current, temperature, etc.

4.5 XADC: Analog to Digital Converter

The XADC has two ADC with 12-bit precision, 1 Mega sample per second (MSPS) and on-chip sensors. The ADCs and sensors are fully tested and provide a general-purpose, high-precision analog interface for a range of applications. Fig. 4.3 shows a block diagram of the XADC. The ADCs support a range of operating modes like externally triggered and simultaneous sampling on both ADCs and various analog input signal types, for example, unipolar and differential. The ADCs can individually multiplex up to 17 external analog input channels.

The XADC also includes sensors like on-chip die temperature sensors and on-chip power supply voltages measurement. The digital data generated after conversion of analog voltage is stored in some fixed registers called status registers. These registers can be accessed through the FPGA interconnect using a 16-bit synchronous read and write port which is known as the dynamic reconfiguration port (DRP). ADC conversion data can also be accessed through the JTAG TAP. For JTAG TAP, users are not required to instantiate the XADC because it is a dedicated interface that uses the existing FPGA JTAG infrastructure. As discussed later, if the XADC is not instantiated in design, the device operates in the default mode that monitors on-chip temperature and supply voltages.

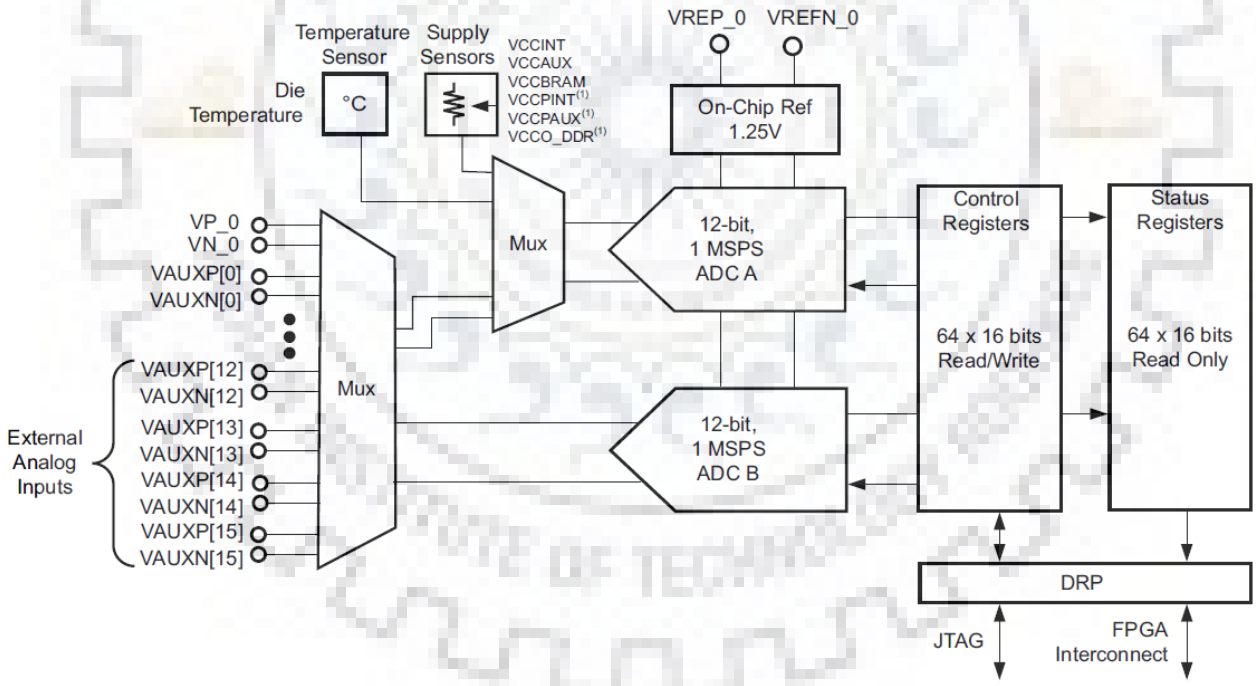


Fig. 4.3 XADC Block Diagram.

4.6 Interrupt Controller

An interrupt is used to pause the normal program execution sequence of KCPSM6 for time being and execute some high priority task. This means that when the 'interrupt' input is driven High ('1'), the interrupt controller will force KCPSM6 to abandon the code that it is executing, save its current operational state and divert its attention to executing a special section of program code known as an Interrupt Service Routine (ISR). Once the interrupt has been serviced, KCPSM6 returns to the program at the point from which it was interrupted and restores the operational states like flag status and register values from the stack so that it can resume execution of the program as if nothing had happened.

The KCPSM6 processor has two pins dedicated to interrupts; an 'interrupt' input and an 'interrupt_ack' output. To initiate an interrupt, the 'interrupt' input must be driven High for 3 or 4 clock cycles. The interrupt input is sampled once every two clock cycles consistent with the instruction execution rate. For this reason, it is vital that the interrupt input is High at the right time to be observed by KCPSM6 and the easiest way to achieve that us to drive the interrupt input High for longer than one clock cycle. There are two fundamental schemes that can be used which can really be described as being 'open-loop' and 'closed loop'. In this design 'closed loop' interrupt mechanism is used which is discussed below –

Closed loop Interrupt

In this scheme, your design drives the interrupt signal High to request an interrupt and then keeps driving it High until KCPSM6 generates an 'interrupt_ack' pulse confirming that it has seen it. This ensures that the interrupt will always be observed by KCPSM6 when it is able to. If interrupts have been temporarily disabled deliberately, or whilst servicing a previous interrupt, then the response will be delayed but the event cannot be missed. Likewise, if KCPSM6 is held in sleep mode when the interrupt is requested it will remain active until KCPSM6 is allowed to wake up and observe it.

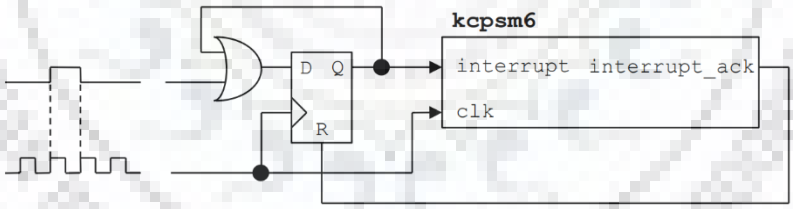


Fig. 4.4 Closed loop interrupt controller circuit.

Open loop interrupt

The simplest way of initiating an interrupt is to generate an active High pulse that has a duration of 2 clock cycles. The pulse can be longer but should have returned Low before the ISR completes otherwise KCPSM6 will immediately think there is another interrupt to service. Once KCPSM6 observes the High level on its interrupt input it will abandon the next instruction and immediately move to the ISR.

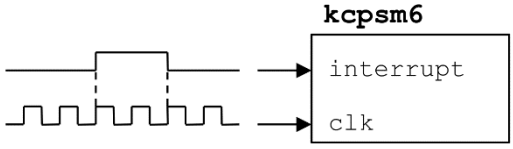


Fig. 4.5 Open loop interrupt controller circuit.

The simplicity of the 'open-loop' method is obvious but it must also be recognized that any open loop system has its limitations. In this case, there are chances that KCPSM6 will miss an interrupt request and therefore fail to service it. This could happen if the interrupt controller is already servicing a previous ISR or the main program has deliberately disabled interrupts.

4.7 Deciding Logic

The deciding logic block whenever called upon by the interrupt controller sends a signal to XADC to read the die temperature voltage and convert it to readable temperature and depending upon the temperature decides the value of the voltage that has to be set.

Die Temperature [°C]	Corresponding voltage value [V]
27	1.1
28	1.08
29	1.06
30	1.04
31	1.02
32	1.0
33	0.98
34	0.96
35	0.94
36	0.92
37	0.90
38	0.88
39	0.86
40	0.85

Table 4.2 Corresponding voltage value for each temperature

The flow chart of this deciding logic is shown in Fig. 4.5. Initially, the main menu is being displayed and actions are performed depending upon input based on the main menu. An interrupt signal is generated every 10 seconds from the main source file which is received by the interrupt controller and. Whenever a valid interrupt occurs deciding logic reads analog data from die temperature sensor through XADC and convert it to a valid temperature value. Depending upon the temperature of die an increment or decrement signal is sent to the PMBus module which in turn changes the supply voltage of FPGA core by 20 mV/°. As the supply of core is decreased for increased temperature the resultant temperature is decreased and the core is settled at an intermediate temperature and supply voltage saving power and controlling die temperature. Table 4.2 states the corresponding value of FPGA core voltage varying from 1.1 V to 0.85 V for a corresponding change in temperature from 27 °C to 40 °C respectively.

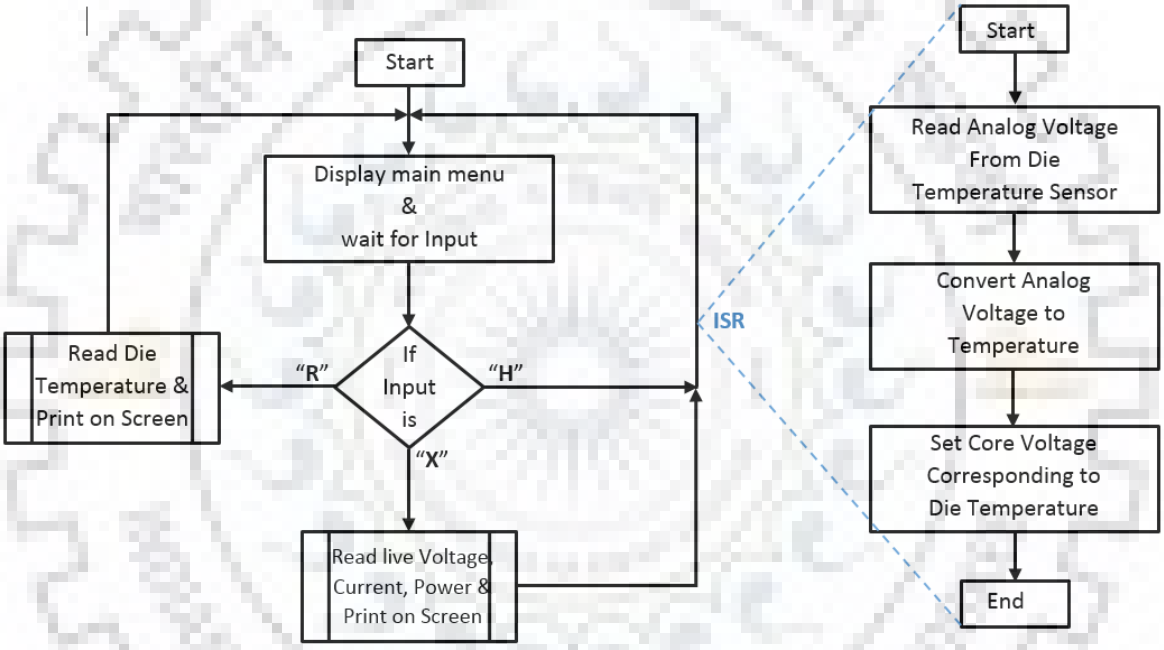


Fig. 4.6 Flow chart of the functioning of the DVS circuit.

4.8 Summary

This hardware level DVS algorithm was verified in the Xilinx Virtex-7 series VC707 board. The assembly code containing XADC and interrupt handler code was compiled using KCPSM6 assembler and converted to Verilog file. 500 1024-bit ripple carry adders were used as test circuit to increase the temperature of the die, the adders were controlled by five external switches for 100 adders per switch. Cutecom serial terminal was used to read output from UART of the board which was die temperature and die supply voltage at an interval of ten seconds. Initially, when the DVS algorithm was switched off and adders were switched on, die temperature was seen increasing till 38 °C and then settling there. But when DVS algorithm was enabled and then adders were switched on, initially temperature increased till 38 °C but later it was settled on 34 °C and die voltage was seen changing for each degree change in temperature as per Table 4.2 above.

This was a very basic implementation of DVS, as it was governed by a LUT but this can be further extended by dynamically increasing or decreasing supply voltage by the variable amount for some change in temperature depending upon the activity on the die. But again that becomes a tedious task of OS, there are certain limitations to implementing DVS on hardware level only. Fig. 4.7 shows a picture of UART data received from the board and options available to be performed.

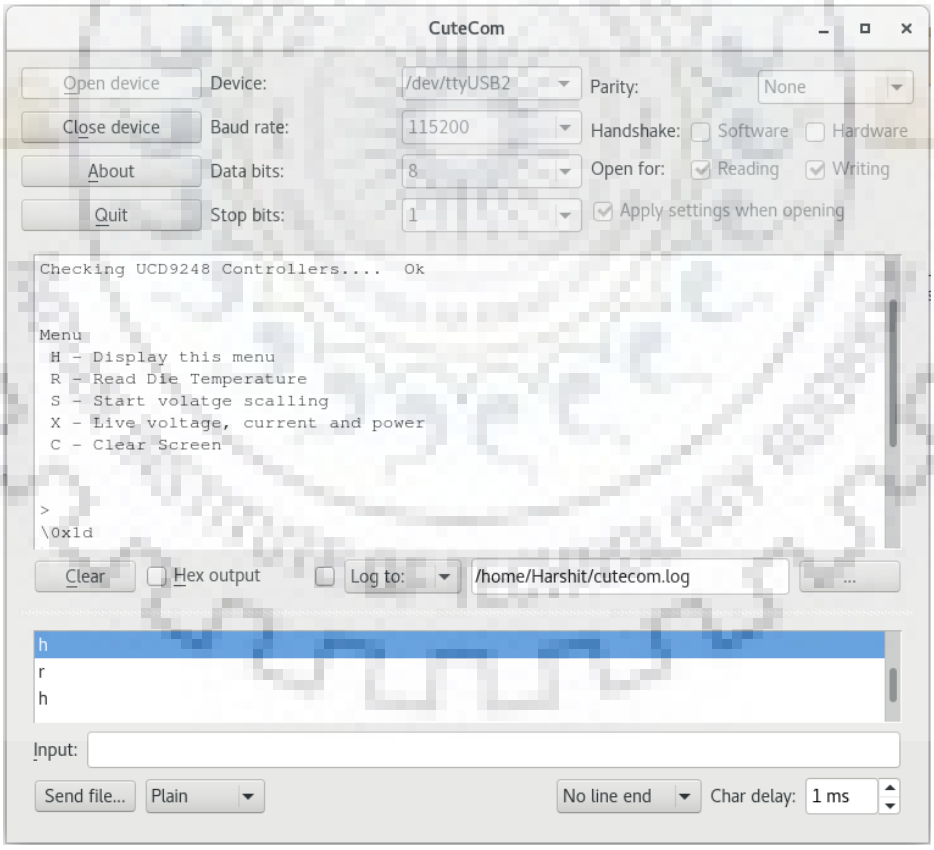


Fig. 4.7 Cutecom serial terminal reading UART data from VC707 board.

Chapter 5: Conclusion

A self-referenced temperature sensor with configurable delay cells is proposed. Single delay cell is used in the design to obtain two different frequencies. Frequency ratio is used to convert temperature information into digital bits, which has the additive advantage of removing temperature nonlinearities due to the impact of mobility. Different temperature sensitivities are obtained by creating the threshold voltage difference between the transistors used in the oscillator. Reverse short channel effect was leveraged to create the threshold voltage difference between transistors. Area and power overhead of voltage regulators and reference clock were avoided by making the design insensitive to supply voltage. Supply sensitivity of the design was reduced by balancing the junction capacitance of the inverter stage. The design was simulated post layout in 65 nm CMOS technology and occupies an active area of 0.00053 mm^2. Based on the post-layout simulation results on both the approaches (this and [4]), we can draw a comparison for a few key features including temperature sensitivity, supply sensitivity, area, power consumption, and most importantly mismatch variations. Process and mismatch variation is the base of our approach because the previous model [4] suggested that mobility effect will be nullified because of frequency division but that was only theoretically possible. With the Monte Carlo simulations, we proved that mobility is going to play a significant role and introduced an error of ± 4.43 °C while our proposed model tries to reduce this effect, and was quite successful in doing so since Monte Carlo simulations in our approach show an error of ± 2.6 °C. The supply sensitivity is found to be 0.032 °C /mV having a peak-to-peak nonlinearity of 2.87 °C with two-point calibration and no polynomial correction for a temperature range of 0-100 °C. In addition, the digital circuitry used to divide frequencies has to be changed in our approach which will reduce hardware significantly.

References

[1] Ronen, Ronny, et al. "Coming challenges in microarchitecture and architecture." *Proceedings of the IEEE* 89.3 (2001): 325-340.

[2] He, Jun, et al. "Highly linear very compact untrimmed on-chip temperature sensor with second and third order temperature compensation." *2010 53rd IEEE International Midwest Symposium on Circuits and Systems*. IEEE, 2010.

[3] Yang, Teng, et al. "Compact and supply-voltage-scalable temperature sensors for dense on-chip thermal monitoring." *IEEE Journal of Solid-State Circuits* 50.11 (2015): 2773-2785.

[4] Anand, Tejasvi, Kofi AA Makinwa, and Pavan Kumar Hanumolu. "A VCO based highly digital temperature sensor with 0.034 C/mV supply sensitivity." *IEEE Journal of Solid-State Circuits* 51.11 (2016): 2651-2663.

[5] Lu, C-Y., and J. M. Sung. "Reverse short-channel effects on threshold voltage in submicrometer salicide devices." *IEEE Electron Device Letters* 10.10 (1989): 446-448.

[6] Jan, M. Rabaey, Chandrakasan Anantha, and Nikolic Borivoje. "Digital Integrated Circuits–A Design Perspective." (2003).

[7] KCPSM6 User Guide, “PicoBlaze for Spartan-6, Virtex-6, 7-Series, Zynq and UltraScale Devices (KCPSM6)” by “Ken Chapman”, Release: 9, 30-09-2014.

[8] KCPSM6 XADC reference design “XADC Interface, Communication & Sampling Conversions” by “Ken Chapman”, 20-03-2013.

[9] “VC707 Evaluation Board for the Virtex-7 FPGA User Guide”, by XILINX, UG885 (v1.8) February 20, 2019.

[10] “PicoBlaze 8-bit Embedded Microcontroller User Guide”, for Extended Spartan®-3 and Virtex®-5 FPGAs Introducing PicoBlaze for Spartan-6, Virtex-6, and 7 Series FPGAs, by XILINX, UG129 June 22, 2011.