Compact Thermal Diffusivity Sensors for On-Chip Thermal Management

Sonmez, Ugur

DOI
10.4233/uuid:915e8fd9-b729-4920-a938-785ecd84e5ee

Publication date
2020

Document Version
Final published version

Citation (APA)
Sonmez, U. (2020). *Compact Thermal Diffusivity Sensors for On-Chip Thermal Management*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:915e8fd9-b729-4920-a938-785ecd84e5ee

Compact Thermal Diffusivity Sensors for On-Chip Thermal Management

Uğur SÖNMEZ

Compact Thermal Diffusivity Sensors for On-Chip Thermal Management

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 12 Maart 2020 om 15:00 uur

door

Uğur SÖNMEZ

Master of Science in Electrical and Electronics Engineering,
Middle East Technical University, Ankara, Turkije
geboren te Istanbul, Turkije

Dit proefschrift is goedgekeurd door de

promotor: prof. dr. K.A.A. Makinwa
copromotor: dr. F. Sebastiano

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. K.A.A. Makinwa,	Technische Universiteit Delft, promotor
Dr. F. Sebastiano,	Technische Universiteit Delft, copromotor
Onafhankelijke leden:	
Prof. dr. ir. W.A. Serdijn,	Technische Universiteit Delft
Prof. dr. ir. B. Nauta,	Universiteit Twente
Dr. J. Shor,	Bar-Ilan Universiteit, Israel
Dr. A. Partridge,	SiTime Corporation
Prof. dr. ir. L.C.N. de Vreede,	Technische Universiteit Delft

TU Delft Delft University of Technology

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Published and distributed by: Ipskamp Drukkers B.V.

ISBN 978-94-028-1971-7

Printed in the Netherlands.

Contents

1 Introduction

1.1. Introduction

Temperature, as an invisible yet palpable physical quantity, plays an important role in our daily lives. Human skin, depending on its anatomical location, contains several tens of temperature sensors per cm^2 [1]. The average household contains a plethora of appliances with temperature sensors, e.g. cooking ovens, coffee makers, refrigerators, kettles, thermostats for heating and clothing irons. These are all designed to operate at specific temperatures, and so temperature sensing errors can cause loss of performance or even device failure.

In some systems, temperature sensors are so critical that a single sensor is not enough. Again, the human skin is a good example: where we feel cold, rather than the sensation itself, is usually relevant. As a system gets more complicated, whether it is biological, electrical, natural or man-made; more temperature sensors are necessary to ensure its performance and reliability. More sensors may also be necessitated by the sheer size of the system, or by speed constraints, e.g. when large temperature gradients must be rapidly detected.

The latter situation is often the case in central processing units (CPUs), systems-on-chip (SoCs) and μprocessors. Thermal management is needed because executing a computationally-intensive process on a CPU can cause local hotspots in a short amount of time. This can compromise reliability, and hence multiple temperature sensors are typically placed on the chip to generate a thermal map [2][3]. A thermal management system continuously monitors this 'thermal map' to distribute the computing load around the chip and guarantee reliability. The finite accuracy of temperature sensors can then affect the system's thermal reliability so much that its performance becomes thermal-management limited [3].

This thesis covers the theoretical foundation, design, and implementation of thermal-diffusivity (TD) based temperature sensors that are intended for thermal management applications in integrated circuits (ICs). Such sensors are ideal for on-chip thermal management because they are accurate, compact and benefit from the scalability inherent to digital CMOS technology. In previous work, such sensors have been used to measure temperature very accurately, with a 3σ inaccuracy down to $\pm$0.2 °C [4]; but these sensors were large (in the 0.1-1 mm^2 range) and were implemented in mature technologies (in 0.16μm CMOS and older). The final result of the research described in this thesis is a TD based temperature sensor implemented in 40nm CMOS, which achieves 0.65 °C accuracy after a single calibration and occupies only 1650 μm^2.

This chapter begins with an overview of the target application: thermal management for CPUs and SoCs, and discusses the requirements on temperature sensors intended for such applications. An overview of CMOS temperature sensors is then given, including an introduction to previous work on TD sensors. It ends with a description of the organization of the rest of this thesis.

1.2. Temperature Sensing for Thermal Management of CPUs and SoCs

Today, microprocessors and other SoCs employ billions of transistors switching at GHz rates. As a result, they can get hot enough to degrade performance and even cause permanent damage. To avoid this, thermal management algorithms, driven by information from on-chip temperature sensors, slow them down or even shut them off when temperatures approach reliability limits. To account for sensor errors, however, such algorithms must incorporate appropriate safety margins. Given that the thermal resistance of a well-designed heat sink may be as low as 0.5 °C/W, a 5-°C margin corresponds to 10 W of unused power [2]. Since a typical microprocessor dissipates slightly less than 100 W, the 10W margin due to sensor inaccuracy represents a significant loss of computing performance, and thus motivates the design of accurate temperature sensors.

In multi-core microprocessors, substantial thermal gradients and hot spots may also occur, whose location is a dynamic function of workload. Thus, multiple on-chip temperature sensors are required, both to ensure reliability and to optimally spread the workload over different cores [5][6]. Since the location of hot spots cannot be easily predicted at design time, on-chip sensors must be small enough to be deployed in large numbers (up to 44 in modern microprocessors [7]), and for their position in the layout to be flexibly moved, even at a late stage of development [5][6].

Accuracy requirements must be satisfied while minimizing calibration effort, which could otherwise significantly increase manufacturing costs, especially when tens of sensors per chip are involved. The toughest requirements are around the reliability limit, with typical specifications being ±1 °C at 70 °C, and only ±3 °C at 50 °C [5]. Moreover, to accurately detect thermal transients with slopes as high as 0.5 °C/ms [5][6], sensor resolution must be less than 0.5 °C, even with measurement times as short as 1 ms.

Moreover, such temperature sensors must be able to co-exist with high-density digital circuits operating at GHz-rate clock speeds. This requirement means that they must operate from noisy digital supplies, which are typically lower than 1V [8]. Therefore, a good DC and AC power supply rejection ratio (PSRR) is necessary. It also means that temperature sensors with analog-output are undesirable, since the presence of high-frequency digital noise makes it very challenging to transmit analog currents or voltages across the chip without picking up interference. To avoid this problem, all temperature information from the sensor should preferably be transmitted digitally.

All these requirements on accuracy, speed, area, ease of calibration, PSRR, and availability of digital output make temperature sensor design challenging. However, several architectures have been described in the literature to meet these requirements and enable dense on-chip thermal monitoring. They can be split into three broad categories:

1. Compact digital-output temperature sensors that can meet the accuracy, PSRR and area requirements [6][5][9].

2. Networks of large, but accurate, absolute temperature sensors in combination with inaccurate, but small, relative sensors to measure the temperature deviation of hot spots compared to the average die temperature [10].

3. Small, analog sensor ’blocks’ distributed across the chip, and using a central readout to multiplex and then digitize their outputs [8][11].

Even though sensor elements such as diodes and MOS devices can be used as tiny temperature sensors (approach 3), guaranteeing the quality of their analog output signals in a hostile SoC environment is a challenging task. Although they can be shielded, this is costly in terms of area, especially if multiple (10+) analog sensors are to be used on the same die.

Using a network of absolute and relative sensors (approach 2) works well when the objective is to make a thermal map of the SoC [10], but its usefulness is limited in situations where an accurate measurement of a particular hot-spot is needed. This is because accurate (absolute) temperature sensors are relatively large, which makes them difficult to locate flexibly in the layout.

For these reasons, this thesis will focus on the first approach: using temperature sensors that are small, fast, accurate and have digital input/outputs. Such sensors can operate in a hostile SoC environment without adding a burden to chip floor-planning. Table 1.1 shows the desired specifications of a temperature sensor for thermal management applications. Of particular note is the area requirement of $<10000\ \mu\text{m}^2$, which is 10-100x smaller than that of general purpose temperature sensors [12][13].

Another important specification is the number of temperature calibration points that should be used to achieve the target inaccuracy at the throttle temperature. Calibration at multiple temperatures means that sensor non-linearity can be better characterized and removed in post-processing. However, this means that every sensor must be tested at multiple temperature points, which takes a long time due to temperature stabilization requirements. This, in turn, increases SoC characterization time and cost. Therefore, multi-temperature characterization is expensive

Table 1.1: Table showing performance of state-of-the-art temperature sensors for thermal management applications

	Target Specs for Thermal Management
Inaccuracy Untrimmed (3σ, °C)	< ± 5
Inaccuracy @ Throttle Temp (3σ, °C)	<± 1
Temp. Range (°C)	0 to 125
Area (μm^2)	< 10000
Resolution (°C, RMS)	< 0.5
Speed (kSa/s)	>1
Supply Voltage (V)	<1
Power (mW)	<10

and highly undesirable. If possible, the trimmed inaccuracy spec in Table 1.1 should be met with the help of calibration at a single temperature.

1.3. Temperature Sensors in Integrated Circuits

Conventional temperature sensors, such as platinum thermistor or thermocouples, have been widely used in automotive, industrial and household applications. CMOS temperature sensors, however, have gradually become more popular. All semiconductor devices and most physical sensors are sensitive to temperature to some degree, so temperature effects must be removed or compensated in demanding applications. Thermal compensation of other devices and sensors is a typical application for CMOS temperature sensors. For example, crystal (XTAL) or MEMS-based frequency references must use temperature compensation to achieve better than 20-50 ppm (parts per million) frequency stability over temperature [14].

A wide variety of devices and methods have been used to measure temperature in CMOS. Sensors have been built by exploiting the temperature-dependent characteristics of diodes [15], bipolar transistors (BJT) [12][16], MOSFETs [17][18], resistors [19][20] or thermal delay lines [4][21]. All of these sensors output an analog quantity, such as voltage, current or frequency. However, the systems that use their outputs, such as CPUs or μprocessors running compensation algorithms, are usually not analog, but digital in nature. Therefore, this analog information must be first conditioned by a readout circuit and then converted to the digital domain by an analog-to-digital converter (ADC). Fig. 1.1 shows the temperature-sensitive element called the 'front-end' or the sensing element, the readout circuitry and the

ADC. Alternatively, in so-called 'smart' temperature sensors, the readout and ADC are merged to directly provide a digital output [12][16][13]. Such systems achieve better power efficiency and accuracy since they remove a part of the circuit which can potentially add noise and degrade accuracy.

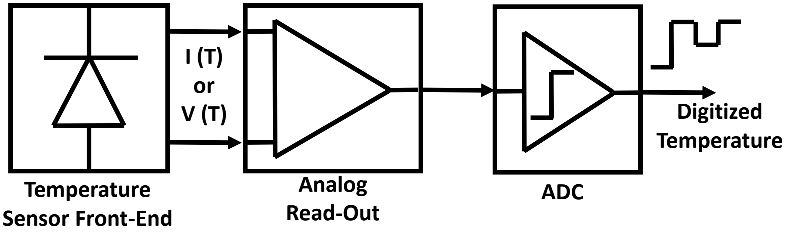


Figure 1.1: Block diagram showing a general-purpose temperature sensor, including the analog front-end element, readout circuitry and the ADC.

Nowadays, 'smart' CMOS temperature sensors can achieve accuracies down to 60 mK [13], resolve temperatures down to 0.1 mK [14], and be as small as 220 μm^2 [15]. The wide variety of flavors in CMOS temperature sensors is too broad to be exhaustively covered in this work, which is why the focus is on the specific application of thermal management. In this section, we will attempt to briefly describe the characteristics of the most popular CMOS temperature sensors used in such applications, starting with BJT-based sensors.

1.3.1. BJT-Based Temperature Sensors

In CMOS technologies, temperature is traditionally sensed by exploiting the temperature-dependency of a BJT's base-emitter voltage (V_{BE}). Under ideal biasing conditions, this is a monotonically decreasing function of temperature, also referred to as being complementary to absolute temperature (CTAT) [12]. V_{BE} can be expressed as:

$$V_{BE} = \frac{kT}{q} ln\left(\frac{I_C}{I_S} + 1\right) \tag{1.1}$$

Here, k is the Boltzmann constant, q is the electron charge, T is absolute temperature in Kelvin, I_C is the collector current and I_S is the saturation current. Due to the strong temperature dependence of I_S, V_{BE} shows a CTAT behavior. Typically, the extrapolated value of V_{BE} is 1.2 V at 0 K (absolute zero) and decreases over temperature with 2 mV/K slope.

Two differently-sized BJT's biased with the same emitter current exhibit a differential ΔV_{BE} voltage. It can be shown that ΔV_{BE} is proportional to T and is a function of the PNP size ratio (p) [16]:

$$\Delta V_{BE} = \frac{kT}{q} ln\left(\frac{pI_C + I_S}{I_C + I_S}\right) \tag{1.2}$$

For $I_C >> I_S$, the expression simplifies to $\Delta V_{BE} = \frac{kTln(p)}{q}$. Therefore, ΔV_{BE} is proportional to absolute temperature (PTAT) and is typically 10s of mV at room temperature. The PTAT behavior of ΔV_{BE} can be combined with the CTAT V_{BE} for a variety of circuit applications. For example, if we sum V_{BE} and ΔV_{BE} scaled by a gain factor α, we can generate a voltage $V_{BG} = \alpha \Delta V_{BE} + V_{BE}$ that is independent of temperature, as shown in Figure 1.2. This is the basic idea behind a bandgap voltage reference. The generated reference voltage is very close to the bandgap of silicon (approx. 1.2V) [12][16].

ΔV_{BE} and V_{BE} can also be combined to sense temperature. From 1.1 and 1.2, we can derive the ratio M as a representation of the absolute temperature:

$$M = \frac{\Delta V_{BE}}{V + V_{BE}} = \frac{kln(p)}{qV_{BG}} T \tag{1.3}$$

The precision of such a sensor depends on the precision of α, ΔV_{BE} and V_{BE}. Fig. 1.2 illustrates the PTAT and CTAT behavior of ΔV_{BE} and V_{BE} respectively, and how they can be combined to generate V_{BG} and measure absolute temperature ratiometrically.

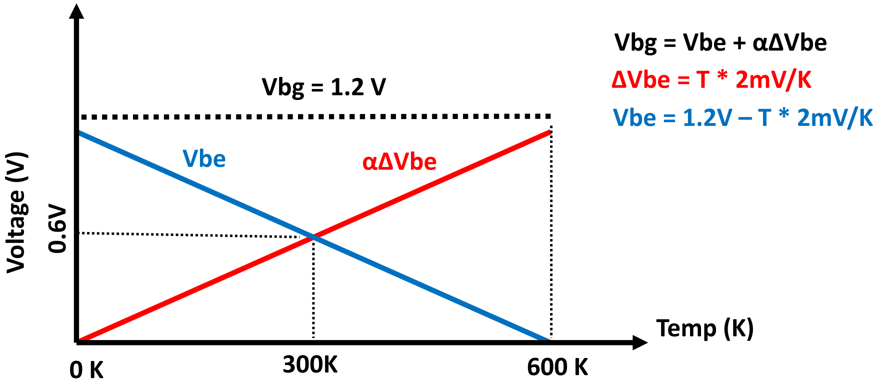


Figure 1.2: Plot showing V_{BE} and $\alpha * \Delta V_{BE}$ and how they can used to generate both a temperature-independent voltage Vbg and also they can be ratiometrically compared to measure absolute temperature

Fig. 1.3 shows a simple temperature sensor that generates ΔV_{BE} and V_{BE}, by using a current mirror circuit to force the same biasing current ($\Delta V_{BE}/R1$) through two differently-sized BJTs with a ratio of p.

The precision of this circuit is limited by the mismatch of the PMOS current mirrors, the offset of the precision amplifier and the spread in V_{BE}. The first two can be improved by chopping or dynamic element matching (DEM) [12] [22] [16]. The latter is a function of process, which can be compensated using a variety of biasing techniques [12]. A positive-feedback current mirror loop [23] can reduce power consumption and circuit area. All in all, the BJT core shown in Fig. 1.3, achieves state of the art temperature-sensing performance in terms of resolution, energy efficiency, and inaccuracy.

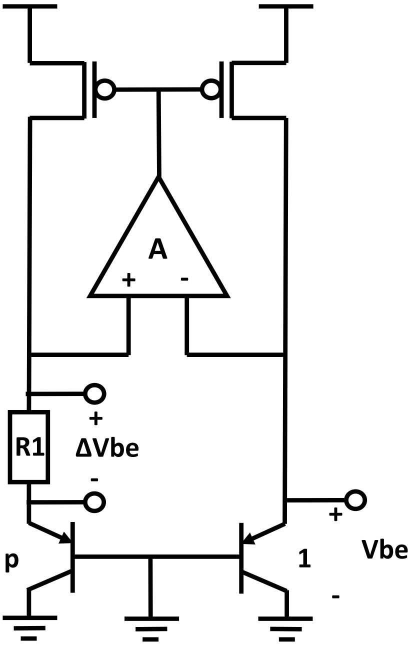


Figure 1.3: Schematic of a BJT core that generates V_{BE} and ΔV_{BE} from a pair of PNPs

A precision temperature sensor also requires an accurate ADC to convert the ratio of ΔV_{BE} and V_{BE} into digital information. Switched-capacitor circuit techniques can achieve this, where ΔV_{BE} and V_{BE} are sampled accurately on capacitors and compared ratiometrically, without using a reference voltage [12][16], to produce the ratio M in equation 1.3. Such a circuit is an example of a smart temperature sensor, as it integrates the ADC into the sensor front-end as closely as possible.

Despite their high energy efficiency and good accuracy, two main factors limit the application of BJTs in thermal management applications [13][16][12]:

1. A relatively high supply voltage requirement, and thus incompatibility with sub-1V operation. This is because V_{BE} is roughly 0.85V at -40 °C.

2. The increase in process spread of BJTs in modern CMOS processes

In recent years, more compact designs intended for thermal management applications [9][24] have also appeared. These designs, usually operating in the current domain rather than directly converting ΔV_{BE} and V_{BE}, simplify the ADC and trade-off either energy efficiency or accuracy for area.

Conversion in the current domain can be achieved, for example, by comparing a PTAT current $\Delta V_{BE}/R_1$ to a CTAT current V_{BE}/R_2. In [9], the difference between these two currents is integrated over a capacitor C, and regulated to zero by a comparator which applies feedback to balance the PTAT and CTAT currents. In [24], the comparator is added inside the biasing loop of an NPN core as shown in Fig. 1.4. Here, the conversion works by using a SAR-algorithm to adjust the value of R2, and hence the CTAT current through it. This current is then compared with the PTAT current through R1. Hence, the loop works as a simple 'temperature comparator'.

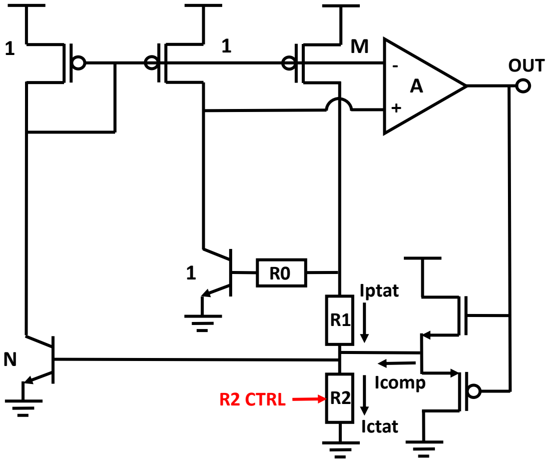


Figure 1.4: Simplified schematic of the BJT-based compact temperature sensor in [24]

In [9] and [24], we can see how recent work has tackled the process spread and large area problems of previous BJT-based temperature sensors by adopting simpler, current-mode techniques. However, they still require supply voltages higher than 1V: 1.8V for [9] and 1.1V for [24]. This is an inherent problem with BJTs and is because V_{BE} can be as large as 0.85V at -40 ^{o}C. Combined with the headroom requirement of the PMOS current mirror, the minimum supply must be equal to or above 1V for most applications. In a modern CPU/SoC; however, sub-1V supply voltages are frequently used.

1.3.2. MOSFET-Based Temperature Sensors

MOSFET-based (or MOS-based) temperature sensors are the designs of choice for sub-1V applications [23][25][8][17][26][27][18]. This is because the threshold voltage of a MOSFET is typically lower than V_{BE} of a BJT. Moreover, these sensors have received considerable attention since their performance inherently scales with process. Furthermore, their device parameters are tightly controlled in all modern CMOS processes.

Various parameters of a MOS device (its transconductance, threshold voltage, etc.) are temperature dependent, and so various classes of MOS-based temperature sensors have been developed. These can be grouped into:

- Dynamic-Threshold or DTMOS-based sensors, which exhibit an exponential trans-conductance and so behave like BJTs. They can then replace BJTs in temperature sensors [23]. The voltage and temperature sensitivity of a DT-MOS is roughly half that of a BJT. As such, it can work with sub-1V supplies at the expense of 2x worse resolution and accuracy when compared to BJT-based sensors [23].

- Subthreshold-based sensors, where the temperature dependent gate-to-source voltage of a MOS device in strong sub-threshold (or weak inversion region) is measured [25][8].

- Delay or VCO-based based sensors, where the temperature dependence of MOS device delay is measured [17][26][27]. This delay is usually a complicated function of transconductance (g_M) and the threshold voltage of the transistor. Despite this complexity, delay-based sensors are popular since their outputs are already in the digital domain.

MOS-based temperature sensors are generally quite compact, achieving areas down to 1000 μm^2 area [18]. This is especially true of delay-line or ring-VCO based architectures since both delay lines and VCOs can be quite small: in the range of hundreds μm^2.

Despite their small area and simplicity, delay-based sensors also have two potential drawbacks: supply sensitivity and a multi-point trimming requirement. The first is due to the fact that MOS or CMOS-based delay lines are generally quite sensitive to supply voltage variations [17]. This problem can be alleviated by comparing the output frequencies of different oscillators, i.e. integrated crystal and reference RC oscillators, and/or by using native I/O devices to improve circuit PSRR [25]. Despite these improvements, the design exhibits a supply sensitivity of 1.3 °C/V.

Most MOS-based sensors have to be trimmed at multiple temperature points to remove second-order effects and obtain a linear characteristic over temperature [17][25][18]. For example, for delay based sensors, the variation in parasitic capacitors over multiple devices can cause significant errors over temperature. This requirement for additional trim points increases the cost of such sensors, which is especially relevant for thermal management applications.

DTMOS-based sensors, being fundamentally similar to BJT-based sensors, do not suffer from the multi-point trim problem [23][28]. Up to now, only one small (0.02mm^2) DTMOS based temperature sensor has been reported in nanometer CMOS [28]. However, there is no fundamental reason why DTMOS designs cannot be made smaller, as they can be directly substituted into a compact BJT-based front-end.

1.3.3. Resistor-Based Temperature Sensors

Resistor-based sensors measure temperature by comparing the resistance of a temperature-sensitive resistor to that of a temperature-insensitive one. Traditionally, a Wheatstone bridge is used to facilitate such a comparison. This requires at least four resistors, which must be large to guarantee good matching. Moreover, the bridge output is in the analog (voltage) domain, and a precision ADC is required. Since the temperature coefficients of typical resistors are strongly non-linear, single-temperature trim is generally insufficient, and two-temperature trim should be used along with systematic non-linearity correction [29]. Due to these reasons, resistor-based temperature sensors have traditionally been large and not suitable for thermal management applications.

Resistors can also be combined with capacitors to build RC-delay or RC-filter based temperature sensors. In [20], temperature-sensitive resistors are used to build a Wien-bridge bandpass filter, whose phase shift is then a function of temperature. A phase-domain readout measures this phase shift, from which the temperature information can be extracted. It achieves a resolution of 3 mK in a 32 ms conversion time and consumes only 31 μW. However, its active area (90000 μm^2) is quite large. In [30], a smaller 6800 μm^2 sensor based on the RC-delay is demonstrated; but it requires trimming at 2 temperatures.

In recent years, there has been renewed interest in adopting resistor-based temperature sensors for thermal management applications. This arises from the following advantages of resistors: no voltage headroom limit, excellent resolution, and energy efficiency figure-of-merit (FoM), and being commonly-used in CMOS processes, well-documented aging and long-term stability characteristics. Furthermore, unlike MOS devices or BJTs, the behavior of resistors is similar in both FinFET and planar CMOS technologies.

Resistor-based temperature sensors have been shown to work with a supply voltage down to 0.7V [31] in a FinFET technology. They can also achieve excellent energy efficieny, with a resolution FoM of 32fJ/K^2 [19]. Low supply sensitivity is also possible, with [32] reporting only 0.23 °C/V. In [33], a compact 7000 μm^2 resistor-based temperature sensor in 65nm CMOS is presented. Just like a Wien bridge architecture, it combines a temperature-sensitive resistor with capacitors to build a fully-differential poly-phase filter. Then, via a frequency-locked loop (FLL), a ring-VCO is locked to the temperature-dependent delay of this poly-phase filter. After two-temperature trim, the 3σ inaccuracy is only $\pm$0.15°C. Despite this excellent performance, an area of 7000 μm^2 can make it challenging to place it close to the SoC hot spots.

Thus, resistor-based sensors are good candidates for thermal management applications where sensor areas in the range of 5000 to 10000 μm^2 are acceptable. There is no fundamental reason why even smaller sensors cannot be implemented; however, none has been published so far.

1.4. Thermal Diffusivity (TD) Based Temperature Sensors

Another option for measuring temperature is to use the temperature-dependent thermal diffusivity of silicon. Thermal diffusivity (TD) describes how fast 'heat signals' travel through a volume of silicon. For silicon itself, it turns out TD is a well-defined function of absolute temperature and has an approximately $T^{-1.8}$ behavior [34], where T is the absolute temperature in Kelvins. This high sensitivity makes TD attractive for absolute temperature measurements.

TD is a mechanical, rather than electrical, property of silicon, and this gives it two significant advantages for use in thermal monitoring applications: it is relatively immune to process variations, and there are no voltage headroom requirements that necessitate a specific supply voltage for the sensor to work. Process variations in modern CMOS are mostly due to doping and lithographic errors, and TD is naturally resistant to any change that does not mechanically alter the silicon lattice. This has been documented in [34], where it is shown that the presence or absence of an n-well around a TD sensor has a negligible effect on its accuracy. This leaves lithography related errors as the dominant source of inaccuracy of TD sensors. This is because TD is usually determined by the time it takes for heat waves to travel in silicon over a well-defined distance [34]. Any inaccuracy in this distance, caused by imperfect lithography, results in an error of TD measurement.

One interesting feature of TD is its compatibility with CMOS process scaling. As lithography techniques improve with scaling, geometries are defined more accurately on silicon, which also enhances the precision of TD-sensors. This makes TD sensors well suited for use in nanometer CMOS processes, in which high-power SoC and CPUs are typically implemented. Scaling of TD-sensors with lithography and its potential limitations are discussed in more detail in section 2.6.1.

One way of measuring the TD of silicon is shown in Fig. 1.5. First, a heater converts an electrical pulse into a heat pulse which is typically implemented by a diffusion resistor. The heat pulse diffuses through the silicon and is detected at a fixed distance by a heat detector, which is typically a relative temperature sensor, and is converted back into the electrical domain.

Since we are interested in thermal 'speed', we can do a time-domain measurement. This can be done by measuring the time difference between the two electrical-to-thermal domain conversions: first from electrical to thermal, and then back again. The mechanical structure that operates the conversion(s) between the electrical and thermal domain is then called an 'electro-thermal filter', or an ETF. ETF design and behavior is explained in detail in Chapter 2 of this thesis.

Relying on the thermal-mechanical property of silicon makes an ETF resistant to process spread, but this comes at the expense of power consumption and energy efficiency. Electrical-to-thermal conversions are lossy in silicon: most of the heat dissipated by the heater will be lost into the silicon substrate, and the resulting temperature variations are quite small: typically <1 °C (RMS) for 1-10mW power dissipated in the heater. This causes the output to be very 'noisy', as the sensor signal is weak, but there is ample thermal noise in the environment. Heater power is

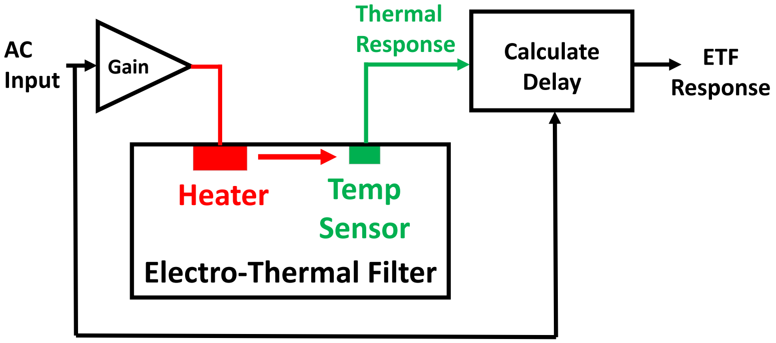


Figure 1.5: Simplified block diagram of a sensor measuring silicon's thermal diffusivity via delay

typically increased to the mW level to solve this problem, but this means that ETFs are an order of magnitude less energy-efficient than, for example, BJT sensors [12][35]. The primary challenge of ETF design is thus improving their SNR without compromising on accuracy. For thermal management applications, the increase in power is tolerable since SoCs typically consume 10s or even 100s of Watts.

All existing TD sensors have used ETFs in combination with various time-domain readouts. Most readouts have been based on a phase-domain sigma-delta modulator (PDΣΔM) [4][21], or a frequency-locked loop [36]. The aforementioned ETF and PDΣΔM combination, shown in Fig. 1.6, works as follows: A frequency reference is used to drive the ETF's heater, thus creating a delayed signal (at the same frequency) at the ETF's output. This signal is then mixed with one of multiple delayed version(s) of the original reference, where the exact amount of delay is under the modulator's feedback control. In such a loop, feedback happens in the phase domain, and hence the loop locks to the condition where the feedback phase is in quadrature with the ETF signal. In the figure, the signal driving the ETF is called F_{DRIVE}, while the demodulating signal (at the same frequency) is called F_{DEM}. The phase shift between the two signals is set by a phase DAC, with a phase shift of Φ_{DAC}. PDΣΔM operation is discussed in more detail in Chapter 3.

The frequency-locked loop (FLL) architecture, shown in Fig. 1.7, resembles a simplified type-I phase-locked-loop (PLL). Rather than relying on the precision of a reference clock, the FLL architecture uses a voltage controlled oscillator (VCO) and locks the VCO frequency to the ETF's thermal delay. Similar to the PDΣΔM, a mixer is used to facilitate this locking behavior. The original and thermally-delayed versions of the VCO signal are mixed, and the DC error is integrated. The loop is only DC-stable when the ETF's phase shift is 90^o or $\pi/2$, and therefore the VCO is locked to the corresponding frequency.

Even though the FLL architecture does not need a precision clock source, it

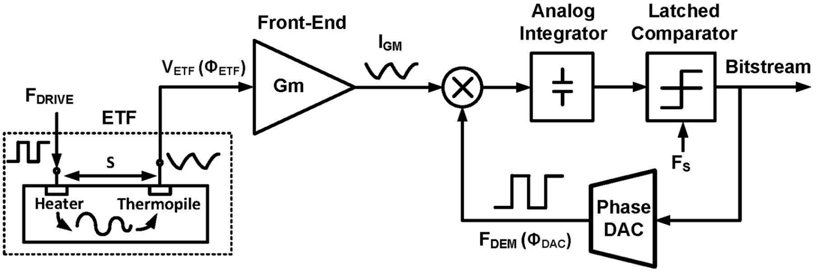


Figure 1.6: Block diagram of an ETF combined with PDΣΔM to read out the thermal delay inside the ETF

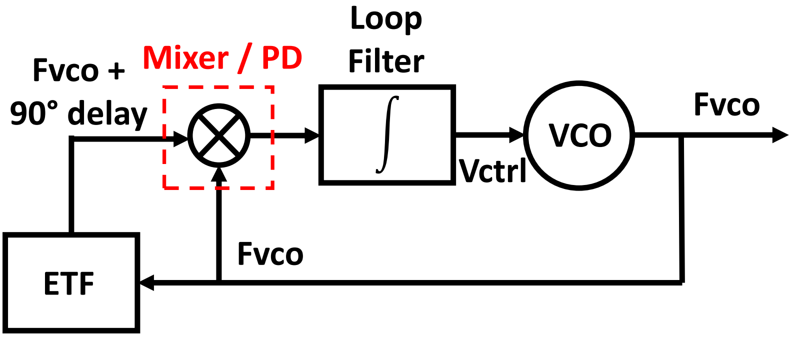


Figure 1.7: Block diagram of a FLL architecture combined with an ETF

has one disadvantage that makes it undesirable for use in precision TD-based temperature sensors: It needs a high-frequency counter to digitize its temperature information, whose power consumption can be quite significant [22]. PDΣΔMs only need decimation filters, which typically operate at low-frequencies.

For this reason, this thesis will focus on PDΣΔMs rather than FLLs as the architecture of choice. Chapter 3 describes in detail how PDΣΔMs, and especially a digital-friendly version realized with a VCO and up/down counters, are a natural fit to the requirements of thermal management applications.

1.4.1. Prior Art on TD Based Sensors

In prior work, TD sensors have been designed for two applications: temperature sensing [4][37] and frequency references [21][36]. In the latter application, ETFs are typically used as timing references, in conjunction with another precision temperature sensor [21], or a temperature-insensitive ETF [36] to compensate their temperature dependence. As frequency references, ETFs can achieve ±1000ppm (3σ) accuracy after a room temperature trim. This specific application is beyond the scope of this work, but similar architectures and circuits are employed in TD frequency reference and temperature sensors.

As temperature sensors, ETFs have performed well in low-cost precision applications, where good absolute temperature accuracy is required without costly trimming procedures. The design in [4] achieved ±0.2°C (3σ) untrimmed inaccuracy from -55 to 125 °C, rivaling state-of-the-art BJT based temperature sensors. By using a temperature-insensitive silicon oxide ETF in combination with a temperature-sensitive silicon ETF in an SOI process, and ±0.4°C (3σ) untrimmed inaccuracy can be achieved without using a precision frequency reference [38].

Prior TD-based temperature sensors have targeted high-precision applications and so are not suitable for thermal management since they are too large (>0.1 mm^2 area), too slow (<10Sa/s sampling speed), require high-voltage supplies and were implemented in older (0.16μm or above) technologies. The design in [4], for example, occupies an area of 0.18 mm^2, consumes a total of 3mW, and achieves a resolution of 30mK at a conversion rate of only 0.32 Sa/s. To facilitate SoC thermal management, this sensor needs to be at least 10x smaller, and 1000x faster, ideally without burning more power. Resolution can be relaxed by 6-10x, which allows some speed improvement (36-100x) for the same energy efficiency, but this is not enough to meet the necessary (>1 kSa/s) sampling rates.

The goal of this thesis is to implement a TD-based temperature sensor that overcomes these limitations and so can be used as a compact, accurate sub-1V temperature sensor in a nanometer CMOS process. Such a sensor would demonstrate the feasibility of TD-based temperature sensing in thermal management applications and unlock more potential applications where TD can be used. In the end, this goal is met by the final design: Two sensors in 40nm CMOS that occupy only 1650μm^2, achieve down to 0.24°C (RMS) resolution at 1kSa/s rate and inaccuracy of 0.65 °C (3σ) after a room-temperature trim. The sensor consumes 2.5mW from a 0.9-1.2V supply, achieving the sub-1V operation goal without consuming more power than the prior art. It is also the first ETF to be realized in nanometer CMOS, and it demonstrates the benefits of scaling for ETFs.

1.5. Organization of This Thesis

The next chapter discusses the design of compact and energy-efficient ETFs. It begins with an extensive analysis of heat transport in silicon, especially for short-distances (<10μm) as is the case in energy-efficient ETFs. Practical limitations to ETF inaccuracy, such as lithography and self-heating, are then discussed; and this knowledge is used to develop a model for predicting ETF accuracy and resolution. To conclude the chapter, two novel ETF designs are discussed in detail, and their modeled vs. measured performance metrics are compared.

Chapter 3 covers the system-level design of phase-domain readouts that convert ETF phase shift into the digital domain. The chapter starts with a brief discussion on phase-detection in CMOS circuits and then expands into two phase-domain readout architectures: Gm-C integrator based and the novel VCO-based architecture. The advantages and design challenges of the VCO-based architecture are discussed in detail. In chapter 4, the detailed implementation of a TD-based temperature sensor in 0.18μm CMOS is presented. This design, using the Gm-C architecture, is intended to be the first stepping stone towards a scaled design in nanometer CMOS.

Chapter 5 expands on this design and presents the implementation of a TD-based temperature sensor in 40nm CMOS. This design is the first implementation of ETFs and VCO-based $PD\Sigma\Delta M$ s in 40nm CMOS, and the first sub-1V TD-based temperature sensor. Further measurements in the chapter discuss the effect of plastic packaging. Finally, Chapter 6 concludes this thesis. A summary of its novel contributions are made, and a section on future work discusses potential improvements.

References

[1] A. Iggo, Ed., *Somatosensory System. Handbook of Sensory Physiology*. Springer-Verlag Berlin, 1973.

[2] J. S. Lee, K. Skadron, and S. W. Chung, "Predictive Temperature-Aware DVFS," *IEEE Transactions on Computers*, vol. 59, no. 1, pp. 127–133, Jan 2010.

[3] J. Shor and K. Luria, "Evolution of thermal sensors in Intel processors from 90nm to 22nm," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, Nov 2012, pp. 1–5.

[4] C. P. L. van Vroonhoven, D. d'Aquino, and K. A. A. Makinwa, "A thermal-diffusivity-based temperature sensor with an untrimmed inaccuracy of $\pm$0.2 ^{0}c (3s) from -55 ^{0}c to 125 ^{o}c," in *IEEE International Solid-State Circuits Conference*, Feb 2010, pp. 314–315.

[5] J. S. Shor and K. Luria, "Miniaturized BJT-Based Thermal Sensor for Microprocessors in 32- and 22-nm Technologies," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2860–2867, Nov 2013.

[6] T. Oshita, J. Shor, D. E. Duarte, A. Kornfeld, and D. Zilberman, "Compact BJT-Based Thermal Sensor for Processor Applications in a 14 nm tri-Gate CMOS Process," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 3, pp. 799–807, March 2015.

[7] M. F. et. al., "Adaptive energy-management features of the IBM POWER7 chip," *IBM Journal of Research and Development*, vol. 55, no. 3, pp. 8:1–8:18, May 2011.

[8] T. Yang, S. Kim, P. R. Kinget, and M. Seok, "0.6-to-1.0V 279 μm2, 0.92 μw temperature sensor with less than $\pm$3.2/-3.4 ^{o}C error for on-chip dense thermal monitoring," in *IEEE International Solid-State Circuits Conference*, Feb 2014, pp. 282–283.

[9] Y. C. H. et. al., "An 18.75 μw dynamic-distributing-bias temperature sensor with 0.87 ^{o}c(3σ) untrimmed inaccuracy and 0.00946mm2 area," in *IEEE International Solid-State Circuits Conference*, Feb 2017, pp. 102–103.

[10] S. P. et. al., "All-digital hybrid temperature sensor network for dense thermal monitoring," in *IEEE International Solid-State Circuits Conference*, Feb 2013, pp. 260–261.

[11] L. Lu, S. T. Block, D. E. Duarte, and C. Li, "A 0.45-V MOSFETs-Based Temperature Sensor Front-End in 90 nm CMOS with a Noncalibrated $\pm$ 3.5^{o}C 3σ Relative Inaccuracy From -55^{o}C to 105^{o}C," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 11, pp. 771–775, Nov 2013.

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

[13] B. Yousefzadeh, S. H. Shalmany, and K. Makinwa, "A BJT-based temperature-to-digital converter with (3σ) inaccuracy from -70 ^{o}C to 125 ^{o}C in 160nm CMOS," in *2016 IEEE Symposium on VLSI Circuits*, June 2016, pp. 1–2.

[14] M. H. P. et. al., "A Temperature-to-Digital Converter for a MEMS-Based Programmable Oscillator With <$\pm$0.5-ppm Frequency Stability and <1-ps Integrated Jitter," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 276–291, Jan 2013.

[15] G. Chowdhury and A. Hassibi, "An On-Chip CMOS Temperature Sensor Using Self-Discharging P-N Diode in a σ-δ Loop," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. PP, no. 99, pp. 1–10, 2018.

[16] K. Souri, Y. Chae, and K. A. A. Makinwa, "A CMOS Temperature Sensor With a Voltage-Calibrated Inaccuracy of $\pm$ 0.15^{o}C (3σ) From -55^{o}C to 125^{o}C," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 292–301, Jan 2013.

[17] T. Anand, K. A. A. Makinwa, and P. K. Hanumolu, "A VCO Based Highly Digital Temperature Sensor With 0.034 ^{o}c/mv Supply Sensitivity," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2651–2663, Nov 2016.

[18] M. C. et. al., "A 225 μm^2 probe Single-Point Calibration Digital Temperature Sensor Using Body-Bias Adjustment in 28 nm FD-SOI CMOS," *IEEE Solid-State Circuits Letters*, vol. 1, no. 1, pp. 14–17, Jan 2018.

[19] S. Pan and K. A. A. Makinwa, "A 0.25mm2 resistor-based temperature sensor with an inaccuracy of 0.12 ^{o}C (3σ) from -55^{o}C to 125^{o}C and a resolution FOM

of 32fJK^2," in *IEEE International Solid-State Circuits Conference*, Feb 2018, pp. 320–322.

[20] P. Park, K. A. A. Makinwa, and D. Ruffieux, "A resistor-based temperature sensor for a real time clock with ±2ppm frequency stability," in *European Solid-State Circuits Conference (ESSCIRC)*, Sept 2014, pp. 391–394.

[21] S. M. Kashmiri, K. Souri, and K. A. A. Makinwa, "A Scaled Thermal-Diffusivity-Based 16 Mhz Frequency Reference in 0.16 μm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 7, pp. 1535–1545, July 2012.

[22] A. Heidary, G. Wang, K. Makinwa, and G. Meijer, "A BJT-based CMOS temperature sensor with a 3.6pjk^2-resolution FoM," in *IEEE International Solid-State Circuits Conference*, Feb 2014, pp. 224–225.

[23] K. Souri, Y. Chae, F. Thus, and K. Makinwa, "A 0.85V 600nW all-CMOS temperature sensor with an inaccuracy of ±0.4^{o}C (3σ) from -40^{o}C to 125^{o}C," in *IEEE International Solid-State Circuits Conference*, Feb 2014, pp. 222–223.

[24] M. Eberlein and I. Yahav, "A 28nm CMOS ultra-compact thermal sensor in current-mode technique," in *IEEE Symposium on VLSI Circuits*, June 2016, pp. 1–2.

[25] K. Yang, Q. Dong, W. Jung, Y. Zhang, M. Choi, D. Blaauw, and D. Sylvester, "A 0.6nJ -0.22/+0.19 ^{o}C inaccuracy temperature sensor using exponential subthreshold oscillation dependence," in *IEEE International Solid-State Circuits Conference*, Feb 2017, pp. 160–161.

[26] P. Chen, S. C. Chen, Y. S. Shen, and Y. J. Peng, "All-Digital Time-Domain Smart Temperature Sensor With an Inter-Batch Inaccuracy of -0.7 - +0.6 ^{o}C After One-Point Calibration," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 5, pp. 913–920, May 2011.

[27] K. Luria and J. Shor, "Miniaturized CMOS thermal sensor array for temperature gradient measurement in microprocessors," in *IEEE International Symposium on Circuits and Systems Proceedings*, May 2010, pp. 1855–1858.

[28] Y. K. et. al., "A 0.02mm2 embedded temperature sensor with ±2 ^{o}C inaccuracy for self-refresh control in 25nm mobile DRAM," in *European Solid-State Circuits Conference (ESSCIRC)*, Sept 2015, pp. 267–270.

[29] M. Shahmohammadi, K. Souri, and K. A. A. Makinwa, "A resistor-based temperature sensor for mems frequency references," in *2013 Proceedings of the ESSCIRC (ESSCIRC)*, Sep. 2013, pp. 225–228.

[30] J. Angevare and K. A. A. Makinwa, "A 6800-µm2 Resistor-Based Temperature Sensor in 180-nm CMOS," in *2018 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov 2018, pp. 43–46.

[31] J. J. H. et. al., "A 0.7V resistive sensor with temperature/voltage detection function in 16nm FinFET technologies," in *IEEE Symposium on VLSI Circuits*, June 2014, pp. 1–2.

[32] H. Park and J. Kim, "A 0.8-V Resistor-Based Temperature Sensor in 65-nm CMOS With Supply Sensitivity of 0.28 ^{o}C/V," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 3, pp. 906–912, March 2018.

[33] W. Choi, Y. T. Lee, S. Kim, S. Lee, J. Jang, J. Chun, K. A. A. Makinwa, and Y. Chae, "A 0.53pJK^2 7000 μm^2 resistor-based temperature sensor with an inaccuracy of $\pm$0.35^{o}C (3σ) in 65nm CMOS," in *IEEE International Solid-State Circuits Conference*, Feb 2018, pp. 322–324.

[34] C. van Vroonhoven and K. Makinwa, "Thermal Diffusivity Sensors for Wide-Range Temperature Sensing," in *2008 IEEE Sensors*, Oct 2008, pp. 764–767.

[35] U. Sönmez, R. Quan, F. Sebastiano, and K. A. A. Makinwa, "A 0.008-mm2 area-optimized thermal-diffusivity-based temperature sensor in 160-nm CMOS for SoC thermal monitoring," in *European Solid State Circuits Conference*, Sept 2014, pp. 395–398.

[36] L. Pedalà, . Gürleyük, S. Pan, F. Sebastiano, and K. A. A. Makinwa, "A frequency-locked loop based on an oxide electrothermal filter in standard CMOS," in *European Solid State Circuits Conference*, Sept 2017, pp. 7–10.

[37] S. M. Kashmiri, S. Xia, and K. A. A. Makinwa, "A Temperature-to-Digital Converter Based on an Optimized Electrothermal Filter," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 2026–2035, July 2009.

[38] C. van Vroonhoven, D. D'Aquino, and K. Makinwa, "A $\pm$ 0.4 ^{o}c (3σ) -70 to 200 ^{o}c Time-Domain Temperature Sensor Based on Heat Diffusion in Si and SiO2," in *IEEE International Solid-State Circuits Conference*, Feb 2012, pp. 204–206.

2

Thermal Diffusivity in Nanometer CMOS

Temperature sensing in silicon integrated circuits (IC) can be achieved by measuring the temperature-dependent thermal diffusivity (TD) of silicon, a material property that is highly insensitive to doping variations. This chapter describes how TD sensors can be fabricated in standard CMOS technology, and it discusses the limitations of such structures, with a particular focus on their area and scalability.

2.1. Introduction

Traditional temperature sensors in CMOS have relied on the base-to-emitter voltage V_{BE} of BJTs [1] to convert absolute temperature into a voltage that can then be measured accurately. Such sensors can achieve accuracies down to ±60mK after a room temperature calibration [2]. As discussed in Chapter 1, however, the accuracy of BJT-based temperature sensors seems to suffer when they are implemented in nanometer CMOS processes.

As an alternative, the thermal diffusivity (TD) of silicon can be exploited to realize a temperature sensor. This is an attractive approach because TD shows a strong temperature dependence [3] [4] [5] and, being a mechanical property, is insensitive to doping variations [5]. It is a natural choice for temperature sensors that are to be implemented in nanometer CMOS technologies, since it benefits from their ever-improving lithographic accuracy.

This chapter begins by presenting the principles of heat diffusion in section 2.2, with a specific focus on heat transport at small distances (<10 μm). These principles will be used in section 2.3 to study the essential component of TD sensors: the electro-thermal filter (ETF). ETF design in CMOS will be discussed in section 2.4 using three possible ETF geometries as examples.

Section 2.5 introduces a harmonic thermal impedance model for ETFs. The results can be used to estimate an ETF's signal and phase shift over temperature and drive frequency. Two ETFs that have been designed with this model are described in sections 2.7.1 and 2.7.2. Section 2.6 discusses the inaccuracy sources of ETFs: lithography, self-heating, and mechanical stress. Section 2.8 summarizes the accuracy vs. energy efficiency trade-offs present in various stages of ETF design and provides guidelines on how to determine critical design variables such as drive frequency. Finally, the chapter concludes with section 2.9.

2.2. Principles of Heat Diffusion

The heat diffusion equation describes the dynamic distribution of heat in a solid:

$$\frac{\partial T}{\partial t} - \alpha \Delta T = 0 \tag{2.1}$$

Here, T is the absolute temperature, t is time, Δ is the Laplace operator, and α is the solid's thermal diffusivity, which defines the speed of the heat transport. It is one of the most famous differential equations in physics and ties three physical quantities together: temperature, time and thermal diffusivity.

Thermal-diffusivity (TD) based temperature sensing relies on the strong temperature dependence of the thermal diffusivity(α) of bulk silicon [4] [5]. For silicon, α is a mechanical quantity that depends on the various phonon scattering rates in its crystal lattice [3], as well as on absolute temperature. For pure bulk silicon at room temperature (298 K), α is 0.8 cm^2/s and it exhibits an approximately $T^{-1.8}$ behavior up to 1400 K [6]. Above 100 K, α is practically independent of doping levels [7]. Mechanical stress has a small impact on α, and its effect is discussed in section 2.6.3.

Due to the robustness of thermal diffusivity to process variation, a sensor that measures it can also measure temperature very accurately. The use of α to measure temperature parallels the use of silicon's well-known thermal voltage (kT/q) for temperature sensors [8]. However, the experimental results that show the robustness of α [7] [6] have only been conducted with large bulk silicon samples. For a complete analysis, we will need to consider the peculiarities of short-distance thermal transport in silicon.

2.2.1. Thermal Diffusion in Silicon at Short Distances

It has been demonstrated that the heat diffusion equation does not adequately describe heat flow in silicon over short distances ($< 10\ \mu$m), which is often the case in integrated sensors [9][10]. The main contributors to heat flow in silicon are phonons, which are vibrations of the crystal lattice. It is known that phonons of different frequencies (or modes) contribute differently to heat transport [11] and that it is the ensemble of different phonons that together produce the resulting heat transport. Over long distances, all phonons undergo various scattering mechanisms, which slow them down and dissipate their energy [3]. This results in the heat transport behavior described in equation 2.1 and the variable α in the expression takes into account all the scattering events experienced by an ensemble of phonons in the silicon crystal.

As the distance traveled by a phonon becomes shorter; however, there is a chance that a phonon does not undergo scattering. This depends on the phonon's mean free path (MFP), i.e. the mean length of the path that a phonon travels before it undergoes scattering, which is longer for lower-frequency phonons [11]. If the phonon is not scattered, it moves freely inside the crystal lattice, analogous to a particle traveling in free space. Such a phonon contributes differently to heat transport, resulting in what is referred to as ”ballistic phonon transport”. Merely speaking, ballistic transport is a direct transport phenomena, analogous to radiative transport, contrary to equation 2.1 which describes diffusion. Accurate modeling of heat transport in the ballistic phonon regime is outside the scope of this work.

Modeling thermal transport over distances longer than the phonon MFP, but still much shorter than the dimensions of a typical silicon sample, has been a topic of great interest [9] [12]. Experimental results indicate that in silicon, diffusive transport is only valid for distances >10μm [12], while ballistic transport is dominant below 40nm. The region in between can be approximated by an accumulation model, where the phonons are first separated into specific frequencies (in a spectrum) with specific MFPs. Phonons with MFPs smaller than the travel distance are considered to contribute to α, while lower-frequency phonons with MFPs longer than this distance do not contribute at all [13]. This results in a thermal conductivity accumulation function, which describes thermal conductivity as a function of a phonon spectrum with λ denoting the phonon MFPs, and $\lambda*$ denoting the travel distance [13]:

$$k_{accum}(\lambda*) = \int_0^{\lambda*} \frac{1}{3} C(\lambda) v(\lambda) \lambda d\lambda \tag{2.2}$$

Here, $v(\lambda)$ is the speed of a phonon with an MFP of λ, $C(\lambda)$ is the thermal capacitance of a phonon with an MFP of λ, and k_{accum} is the accumulated thermal conductivity. The total thermal diffusivity (α_{accum}) can then be calculated from the relation between thermal diffusivity (α) and thermal conductivity (k):

$$\alpha = \frac{k}{C} \tag{2.3}$$

Here, C is the thermal capacitance of the material and is also known as ρc_p, where ρ is the material density and c_p is the specific heat capacity. For equation 2.2 this definition gives us: $\alpha_{accum} = k_{accum}/C_{accum}$, where C_{accum} (the accumulated or total thermal capacitance) is the integral of $C(\lambda)$.

There have been multiple attempts in the literature to model and understand $v(\lambda)$ analytically, but these are outside the scope of this work. What is relevant is the prediction of these models, i.e. that α decreases as distance reduces, and a smaller portion of the phonon spectrum contributes to diffusion. This means that heat transport approaches a ballistic limit as the travel distance is reduced. If this was not true, and equation 2.1 was correct; then heat would travel too quickly for small distances. For a more detailed treatment, [10] explains how and why this is an unrealistic phenomenon.

Figure 2.1 shows this more graphically. Here, a phonon travels short, medium and long distances (from point A to B, C, and D respectively) in silicon. The mean free path that the phonon undergoes before scattering is shown as the phonon MFP. The two plots on the right show the normalized transient temperature change at the destinations B, C and D; and the speed of the heat transport with respect to inverse distance. The first plot is intended to show the time-of-flight of the phonon for the three cases.

In the third case (D), the distance is long, the phonon scatters, and hence the phonon's time-of-flight is inversely proportional to distance. This means α is fixed. In the second case (C), distance is shorter, and therefore the phonon travels ballistically for the majority of its flight. It arrives at its destination quicker, but not as quickly as expected from diffusion theory. In the first case, the distance is very short, and hence the phonon travels ballistically. Time-of-flight is shortest, but the phonon's speed is bounded. If equation 2.1 were to be used, we would find α to be lower than expected.

This model is consistent with the experimental results obtained in [9] and [12], where heat conduction was found to be a function of the distance of thermal transport. In [12], heat conduction was found to be up to 40% slower than expected for a distance of 1 μm, when compared to a transport distance of 20 μm. A more accurate version of Fig. 2.1 can also be found in [12].

At first, these results seem to contradict the fact that the mean MFP of phonons in silicon is roughly 40 nm; and hence we would expect ballistic effects to be only dominant below 40 nm. The reality is a bit different: phonon frequencies in silicon cover a wide spectrum with MFPs ranging from several nm to 10-20 μm. The distribution of this phonon spectrum is not well understood, especially concerning its contribution to heat transport. However, recent experiments [9][12] have revealed

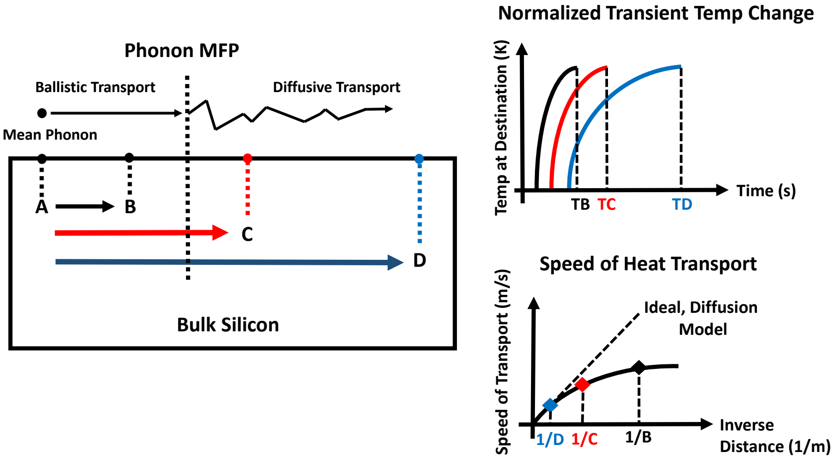


Figure 2.1: Three cases showing the speed of heat transport over distance for a mean phonon. The figure shows the travel path of the phonon over three distances, and the mean free path (MFP) before which the phonon particle travels ballistically.

that lower-frequency phonons with MFPs in the range of 0.1-10 μm mediate the majority of heat transport in silicon. A variety of theories have tried to explain this behavior [11] [14] [15], the theoretical framework for this behavior is still a hot topic.

The experimental results, however, all agree with the accumulation method presented in [13] and equation 2.2, and hence we will use it to express α as a function of $\lambda*$, or distance. Once $\lambda*$ is defined, we can generate a modified α value that can be used in equation 2.1. Experimental results that relate α to distance can be found in [9] and [12]. The important distinction here is to determine when the heat diffusion model should be changed and to what extent. We will call this region of operation the quasi-ballistic operation, where heat transport is both ballistic and diffusive. A simple modification of equation 2.1 for the quasi-ballistic region is to modify the thermal diffusivity according to the distance from the heat source, as presented in [9] and [12].

The dependence of α on distance is expected to be a process-independent mechanical property of silicon. In [9], the effect of doping and temperature on phonon MFP spectra have been characterized. Doping is shown to have a negligible impact, while a 10 % increase is observed between 311 to 417 K, for a spectrum range of 400 nm to 4 μm. There are no results on the effects of mechanical stress, as this has not been well explored in the literature. Despite this, the robustness of α over doping and temperature is encouraging.

With these considerations in mind, a model for thermal diffusivity as a function of distance s was generated from the data presented in [12]. In the cited experiment, data were obtained from the exponential decay of transient heat pulses generated in silicon via laser-induced dynamic grating [16]. Fig. 2.2 shows the normalized model for thermal diffusivity of silicon (α) as a function of distance. Here, the data was generated numerically from the plots in [12] and normalization was done based on the diffusivity of bulk silicon (0.88 cm^2/s) [17]. This data is also tabulated in Appendix A, together with an approximate numerical model determined by a MATLAB extrapolation.

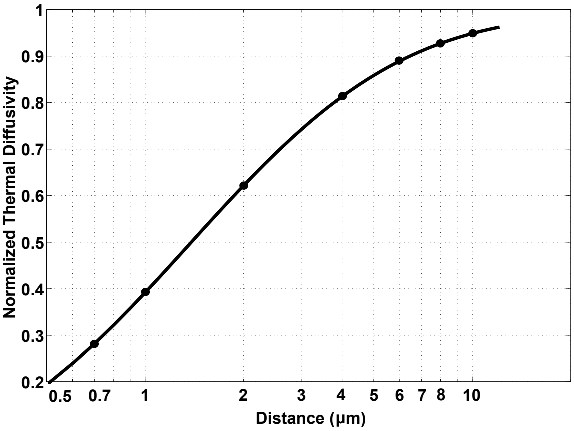


Figure 2.2: Normalized model of α with respect to distance, with normalization reference to the bulk diffusivity of silicon

In the sensors described in this thesis, heat is transported over distances s, ranging between 2 and 10 μm, which is covered by the aforementioned quasi-ballistic model. For distances larger than 10 μ m, the error is less than 5%, and so the heat transport is predominantly diffusive. There is limited data for s<2.4 μm, so the model in Fig. 2.2 was extrapolated to cover distances down to 0.5 μm.

Testing and validating this model would confirm that ballistic thermal transport occurs in bulk silicon. Laser grating experiments require testing to be done in a thin slice of silicon, and thus the results are not directly applicable to bulk models. ETFs, which are built into bulk silicon, are excellent platforms for characterizing the thermal diffusivity of silicon over specified distances.

2.3. Measuring Thermal Diffusivity: Electro-thermal Filters

One way to directly measure α is to inject some heat into silicon (via Joule heating) and observe the temperature at the injection point after a certain amount of time. However, since silicon is a good thermal conductor it is tricky to detect the small temperature rise due to Joule heating in the presence of ambient temperature variations. Since the latter changes very slowly [18], we can separate the two by up-modulating the Joule heat signal to a higher frequency.

A simple structure that generates a high-frequency heat signal is shown in Fig. 2.3. In the figure, the distance s is much smaller than the thickness of the substrate, which can then be regarded as being a semi-infinite volume of bulk silicon. The heat source is embedded at the silicon surface; which is assumed to be covered by a semi-infinite volume of silicon oxide. Silicon oxide has x100 lower α compared to silicon [4][19], and so the heat will mainly flow through the silicon.

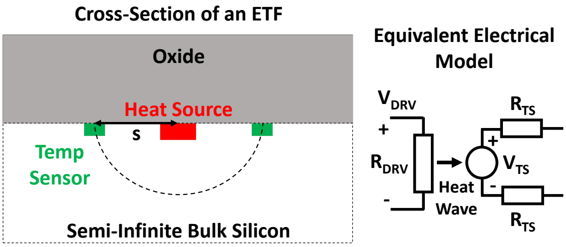


Figure 2.3: Cross section and the equivalent electrical model of a surface heater and temperature sensors placed at a radial distance of *s*. This structure is known as an ETF.

The structure in Fig. 2.3 is known as an electro-thermal filter (ETF), in which an electrical signal is converted to the thermal domain and then back to the electrical domain [20]. The first conversion is achieved by applying an AC drive voltage V_{DRV} over a heater/resistor R_{DRV}. A set of temperature sensors placed at a radial distance from the heater *s* pick up the heat wave, and generate an AC voltage V_{TS}. The output impedance of the sensors is modeled by a resistance R_{TS}. By observing the properties of the heat waveform (amplitude, delay, etc.), we can extract the α of bulk silicon.

The choice of the heater and temperature sensor elements are critical to the ETF design. For the heater, resistors are commonly used [5], but MOSFETs can be used as well [21]. For the temperature sensor, multiple options are available in a CMOS process. The most common are:

- BJTs (or BJT pairs), via the temperature dependence of V_{BE} and ΔV_{BE} [1]
- MOSFETs, via the temperature-dependence of threshold voltage or mobility [22]
- Thermistors, via the temperature dependence of resistance
- Thermocouples, via the Seebeck effect [23]

BJT and MOSFET devices exhibit good signal-to-noise ratio (SNR) and temperature accuracy, but their intrinsic offset means that a complex front-end is required [1]. Thermistors can achieve the best SNR, but they have a large base-line component (offset). The offset intrinsic to thermistors can be much (10x or more) greater than the temperature signal received from the ETF, making it difficult to apply classic offset-reduction techniques such as chopping.

Thermocouples in CMOS have a lower SNR, but they have no intrinsic offset, and several of them can be connected in series (as a thermopile) to get a larger signal [5][23][24]. As a disadvantage, thermocouples can only measure the relative temperature between two points, which means they need a reference temperature point, called the cold junction of the thermocouple. Accordingly, the signal side of the thermopile is called the hot junction. Since the amplitude of the heat waveform decreases exponentially with distance (see equation 2.1), the cold junction can be placed relatively close to the hot junction. Typical values are 0.5-1.5 *s* from the hot junction or 1.5-2.5 *s* from the heater [25].

For these reasons, we will employ thermocouples as the thermal sensing element. The thermocouples will be placed at a fixed radial distance *s* from the heat source. The cold junction is placed much further away from the heater; and to simplify things, we will first assume that it is at room temperature. We can then analyze/design such a structure via equation 2.1. Because of the semi-spherical symmetry in the structure, we can easily solve the equation in semi-spherical coordinates. This comes from the intuition that all the points on a sphere at a distance *s* in Fig. 2.3 are subject to the same heat wave. As mentioned before, we assume that no heat flows through the oxide, since its α is two orders of magnitude lower than that of silicon.

Given all these parameters, the heat diffusion equation can be solved for a simple ETF as in Fig. 2.3. For a periodic, sinusoidal heat signal *H(t)* at a frequency of *F* generated at the point heat source, the temperature phasor at a distance *s* from the heater is given as [27]:

$$T(t) = \frac{H(t)}{2\pi ks} e^{-s\sqrt{\frac{\pi F}{\alpha}}} e^{-js\sqrt{\frac{\pi F}{\alpha}}} \tag{2.4}$$

Here, *k* is the thermal conductivity of silicon as defined in equation 2.3, *s* is the distance, *F* is the excitation frequency, and α is the material thermal diffusivity. In equation 2.4, the temperature phasor T describes the AC or transient behavior of real temperature, similar to a voltage phasor. The ETF can be characterized by a thermal impedance Z, i.e. the ratio of *H* and *T*:

$$Z = \frac{T}{H} = \frac{1}{2\pi ks} e^{-s\sqrt{\frac{\pi F}{\alpha}}} e^{-js\sqrt{\frac{\pi F}{\alpha}}} \tag{2.5}$$

The thermal impedance Z is a direct analog of the voltage impedance of a circuit element. We can imagine heat (H) as a current, and temperature (T) as a voltage potential. Thermal resistance, similar to electrical resistance, is defined as the steady-state ratio of T/H and is proportional to $1/k$. Building on that idea, Z describes both the steady-state and the AC behavior of T(f)/H(f) over frequency.

From equation 2.5, we can extract α by observing the amplitude or phase of Z. α, *k* and *s* will all contribute to the amplitude, while only α and *s* contribute to the phase shift. An amplitude measurement of Z would be further corrupted by the spread in input heat H, and by the offset, gain error and non-linearity of the temperature sensor. Therefore, measuring α by observing the phase is a better choice. Techniques to extract this phase from the output of an ETF are discussed in detail in Chapter 3. In this chapter, we will focus on the intrinsic properties of ETFs, without dealing with readout considerations.

We can derive the phase (Φ) of the ETF in equation 2.5 as:

$$\Phi = -s\sqrt{\frac{\pi F}{\alpha}} \tag{2.6}$$

Since $\alpha \propto T^{-1.8}$ around room temperature, $\Phi \propto T^{0.9}$; and thus the ETF phase is roughly linear with respect to temperature. In practice, this non-linearity can be corrected in the digital domain to achieve a linear relationship between Φ and T [28]. Thus, we can extract temperature information from an ETF with a simple phase-to-digital converter.

2.4. ETF Design in CMOS

In order to build an ETF in a standard CMOS process, two elements are required: a heater and a temperature sensor, usually implemented as a thermopile (a series-connected set of thermocouples). The heater can be a simple diffusion resistor, and thermocouples can be made in CMOS by connecting p+ or n+ active regions and metal layers [5]. A typical CMOS thermocouple (Al/p+ or Al/n+) exhibits a Seebeck coefficient of several 100s of μV/K [29][30] depending on the doping of the active layer. Several of these can be stacked in series to build a thermopile with a sensitivity of several mV/K. Due to the high resistivity of active layers (50-300 Ω/square, depending on the process) such thermopiles exhibit significant thermal noise. In practice, p+ active layers are usually chosen, as they can be placed in an isolating n-well.

Up to this point, we have approximated the cold junction's temperature as being equal to room temperature; but this is unrealistic. Since the cold junction is placed on the same silicon as the hot junction, it will pick up a weaker version of the heat wave; and due to the nature of the Seebeck effect, reduce the voltage output of the thermocouple. As the cold junction is placed farther away, this heat signal will decrease with increasing distance from the heater. In the limit case where it is infinitely far away from the heater, the cold junction's effect becomes negligible. Thus, for a strong ETF signal, we would like the cold junction to be as far away from the heater as possible. This is, however, limited by area and SNR concerns, as will be explained later.

Geometrical placement of the heater and thermocouples is important to maximize the ETF's heat signal while minimizing its thermal noise. Due to power consumption concerns, a typical ETF dissipates 1-5 mW in its heater, while an ETF's thermal impedance (Z) typically varies between 50-500 K/W. Thus, typical AC temperature variations at the hot junction are in the range of 0.1-1 K. Most temperature sensors (including thermocouples) have sensitivities of only a few mV/K, and thus the output voltage of a typical ETF will have an amplitude of at most a few mV. In addition, the resistance of the thermocouple is at least a few kΩs, and thus its thermal noise is significant when compared to the ETF signal. This limits the resolution of a typical compact ETF to 12-14 bits. This is highly dependent on two process related parameters: diffusion-resistor resistivity (assumed as 180 Ω/square); and thermopile (Al/ Si P+) Seebeck coefficient (assumed as 0.25 mV/K).

The same placement of the heater and hot junctions also defines the fundamental accuracy of the ETF. Assume that $s = s_0 + ds$, where s_0 is fixed by design and ds is the random error that exists on s_0. In a CMOS process, geometrical features such as s are defined by the mask dimensions and lithography and ds is expected to be a well-bounded error irrespective of s (see section 2.6.1). Therefore, as s_0 increases by design, the relative error on s and hence its accuracy improves.

These considerations lead to conflicting requirements on the location of the thermopile's hot and cold junctions:

1. For the largest output signal, the thermocouple's hot junctions should be close to the heater, while its cold junctions should be far away

2. For the lowest thermal noise, the thermocouples should be as short as possible

3. For best accuracy, the hot junction should be far away from the heater

These three points summarize the fundamental trade-offs in the design of ETFs. The choice of the distance s is then critical to obtaining the best trade-off between SNR and accuracy. In ETF design, the choice between SNR and accuracy depends primarily on the application. For thermal management applications, resolution and area concerns trump accuracy (to a degree) and hence the choice is for a small s.

The parasitic junction capacitance of the thermocouples constitutes another challenge to ETF design. This capacitance exists between the p+ diffusion resistors and the isolating n-well and is proportional to the thermocouple area. The combination of thermopile resistance R_{TP} and such parasitic capacitance adds further delay to the ETF signal, degrading phase accuracy. This delay, however, can be detected and compensated by an electronic calibration technique introduced in Chapter V, under section 5.3. Nevertheless, the thermocouple area must be minimized to minimize the parasitic RC delay and maximize ETF accuracy.

Various ETF geometries have been implemented in the literature to solve these optimization problems. We will begin with the simplest ETF geometry: the bar ETF.

2.4.1. BAR ETF

Figure 2.4 shows the 2-D layout view of a bar ETF, along with an electrical equivalent model. In this ETF, a long bar-shaped heater is surrounded by sets of thin, long thermocouples perpendicular to the heater. S_{HOT} is defined as the distance from the horizontal mid-point of the bar heater to the hot junction; while S_{COLD} is the distance from the heater to the cold junctions. The hot and cold junctions are defined by the intersection of p+ active regions (green regions) and metal contacts (square white boxes).

For differential operation, the thermocouples are divided into two identical half-thermophiles, each generating half of the ETF voltage V_{ETF}.The common-mode of the ETF is defined by the voltage V_{CM}, which can be freely chosen and only depends on the input specifications of the readout electronics. Each half-thermopile also has an electrical resistance of R_{TP}.

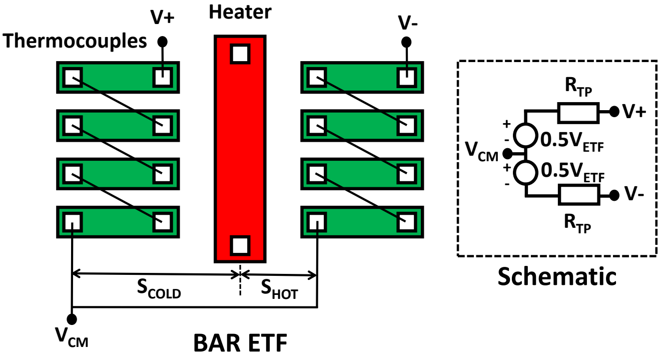


Figure 2.4: 2-D Geometry of a bar ETF and its associated schematic

The dynamic ETF voltage $V_{ETF}(t)$ can be described in terms of the ETF's thermal impedance $T_{ETF}(t)$, Seebeck coefficient γ and number of thermopile arms *N*:

$$V_{ETF}(t) = N * \gamma * T_{ETF}(t) \tag{2.7}$$

Here, $T_{ETF}(t)$ can be calculated by convolving the ETF thermal impedance Z_{ETF} with the heat signal.

The bar ETF is a simple structure that can be built in any CMOS technology, but it's SNR is not optimal [25]. The temperature information at the hot junctions of a bar ETF will only sum up coherently if the phase of the heat wave is the same for all of them. However, this will only happen in the limit case when the heater is much longer than the total width of all thermocouples, or when *s* is much smaller than the thermopile width. Therefore, the heat generated at the edges of the heater, close to its metal contacts, is wasted. This can be seen in [25], where it was shown that a bar ETF's SNR could be improved by 50% if the thermopile's hot junctions are placed on a constant phase shift contour around the heater, as shown in the next section.

2.4.2. Phase Contour ETF

Placing all the hot junctions on a constant phase contour around the heater results in the phase contour ETF, shown in Fig. 2.5. The hot junctions now capture all of the heat generated by the heater in a coherent fashion. Both the hot and cold junctions are then placed on a circular pattern around the heater, where both patterns correspond to a specific phase contour, i.e. the radial distance *s* which defines a specific phase shift. Here, we assume that the heater is small enough to be approximated by a point-source. For small ETFs ($s < 2\ \mu$m), where this assumption may not be realistic, the heater can be laid out in a circular or square fashion to mimic a point source.

In Fig. 2.5, S_{HOT} defines the hot contour and the distant S_{COLD} defines the cold one. This placement guarantees that all hot and cold junctions receive the same heat signal in both amplitude and phase shift; which improves both SNR and accuracy [25].

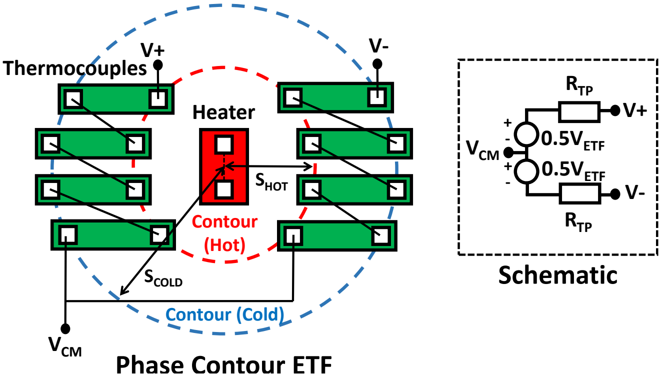


Figure 2.5: 2-D Geometry of a Phase Contour ETF and its associated schematic

We can define a phase contour ETF with the parameters S_{HOT} and S_{COLD}, number of thermophiles *N* and thermopile resistance R_{TP}. Assuming a point heater, a phase contour ETF is radially symmetric, and hence it can be modeled well in radial coordinates with the parameter *s*, as in equations 2.4 to 2.6. Here, *s* can be either S_{HOT} or S_{COLD} to calculate the phase shift and signal amplitude on hot and cold junctions, respectively. For a typical phase-contour ETF and given S_{HOT}, the optimum point of S_{COLD} for best SNR can be mathematically derived as in [25]. The optimum point for S_{COLD} minimizes the signal loss due to the cold junction, as well as the thermal noise due to the thermopiles. The optimum SNR is achieved when the mean length of the thermopiles is about $1.5S_{HOT}$, or when $S_{COLD} = 2.5S_{HOT}$ [25].

In layout, it is hard to define S_{HOT} and S_{COLD} precisely, since the heaters and the hot/cold junctions physically occupy finite area. Therefore, S_{HOT} and S_{COLD} should be considered as the average distances from the center of the heater (approximated as a point source) to the centers of the physical hot/cold junctions. Such junctions are defined by the overlap of metal (Al) and active (P+) layers, and typically occupy areas ranging from 0.01 to 0.04 μm^2.

Phase contour ETFs have been used to build both temperature sensors [25] and frequency references [31]. Despite their excellent accuracy (0.2 ^{o}C inaccuracy for a temperature sensor [32]), phase contour ETFs are rather noisy (12-bit SNR for a 10 Hz bandwidth [31]). This has limited their application to low bandwidth operation, typically less than 1 Hz [32][25][31]. Fortunately, we can improve the SNR of phase contour ETFs by minimizing R_{TP}, and hence thermal noise. This brings us to the newly proposed polygon ETF.

2.4.3. Polygon ETF

As shown in Fig. 2.6, the noise efficiency of a phase contour ETF can be improved by expanding the thermocouples to cover all the area between the hot and cold phase contours. This significantly reduces their resistance (and thermal noise). Note that the figure shows 12 thermopiles for ease of drawing and presentation, but the number of thermopile elements to be included as part of the contour (N) is a design variable. Due to practical and design rule related reasons in layout; an octagonal shape, rather than a non-specific polygon, is commonly implemented. Hence, this ETF geometry is also known as the 'octagonal' ETF.

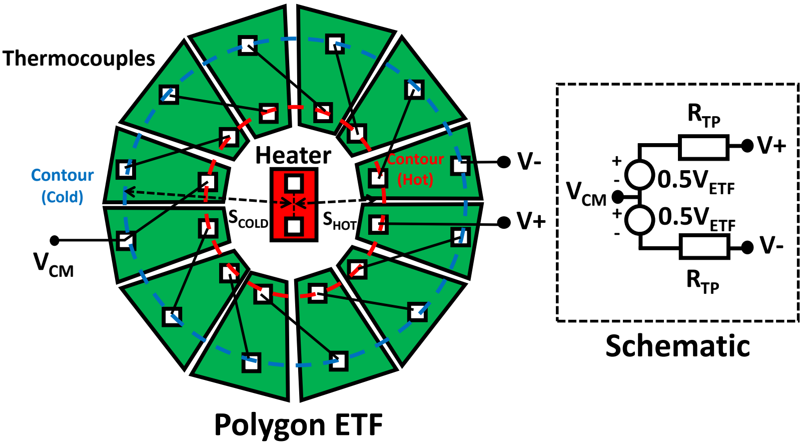


Figure 2.6: 2-D Geometry of a Polygon ETF and its associated schematic

The improvement we gain from adopting the polygon contour geometry is a function of process parameters, such as contact dimensions and thermocouple spacing. Nevertheless, we can make some assumptions to simplify the comparison:

1. Both the polygon and phase contour ETFs have the same number of thermocouples, N.

2. The hot junctions are *s* away from the point heater, while the cold junctions are placed at a distance of (1+X)*s*. This is done for both phase contour and polygon ETFs.

3. The space between the thermocouples is negligible, and the hot junctions are arranged in a circular fashion around the heater.

Due to the first and second assumptions, both phase contour and polygon ETFs will pick up the same heat signal and hence have the same output signal. However, their thermal noise levels will be different due to the difference in their thermopile resistance (R_{TP}). We can calculate the minimum R_{TP} of a phase-contour ETF as:

$$R_{TP}(Contour) > N^2 \frac{R_{SQ}X}{2\pi} \tag{2.8}$$

Here, $(2\pi s)/N$ is the maximum width and Xs is the length of each thermocouple with X as a normalization variable, and R_{SQ} is the thermocouple's resistance per square. The maximum width is calculated with the assumption that N thermopiles fit snugly on the contour at a distance of S_{HOT}, with negligible distance between each thermopile. In practice, the distance is not negligible and their width is always less than $(2\pi s)/N$; which makes R_{TP} always larger than the minimum.

The minimum thermopile resistance of the polygon ETF is:

$$R_{TP}(Polygon) > N^2 \frac{R_{SQ} ln(X+1)}{2\pi} \tag{2.9}$$

The ratio between equations 2.8 and 2.9 is $X/ln(X+1)$. For an SNR optimized phase-contour ETF, X is approximately 1.5, and if we compare an optimized phase-contour ETF to a polygon one, from equations 2.8 and 2.9; R_{TP} is reduced by x2.9 and so the ETF's SNR is improved by 69 %. Due to its superior SNR performance, polygon ETFs will be considered in this work as the ETF of choice.

One drawback of a polygon ETF is a larger parasitic junction capacitance, and the associated phase inaccuracy. A polygon ETF covers a much larger area (typically 2-5x) than a phase contour or bar ETFs, and hence has more junction capacitance. These parasitics, combined with the readout's input capacitance, can be as large as 100s of fF and can introduce a phase shift of several degrees for MHz drive signals. This is less of a problem in thermal management applications, where the accuracy requirements are around $1^o C$. This approximates to roughly $0.1 - 0.3^o$ phase-shift for a typical ETF.

In the measurements outlined in Chapter 5, this phase shift was measured to be 1-1.5 ^{o} for an s = 3.3 μm ETF driven at 1.172 MHz and results in an estimated inaccuracy of 1.5^{o}C (3σ). However, this electrical phase shift can be measured and then canceled in the readout circuitry (see section 5.3 for more details).

While the polygon ETF can be laid out in most advanced CMOS processes, the presence of non-orthogonal angles and incompatibility with strict DRC rules disallows them to be used in FinFET technologies. For FinFET processes, either the BAR or phase-contour ETF geometries should be used.

To evaluate the SNR of various ETF structures and determine the critical distance s of an ETF, a thermal model of polygon and phase-contour ETFs has been developed. The following sections describe this thermal model, starting with special considerations for ETFs with s<10μm.

2.5. A Harmonic Thermal Impedance Model for ETFs

Modeling the sensing element is a critical part of a temperature sensor's design process, and this is no different for ETFs. Before we can begin with readout design, we need to design and determine the ETF's properties. Important properties are output voltage amplitude (RMS), amplitude and phase shift of the heat wave (for a given frequency), RMS thermal noise at the output and the power consumed by the heater. The first two properties are related to the thermopile's Seebeck coefficient γ and the number of thermocouples (N). γ is fixed by the process node, and the relationship between output RMS noise and N is a function of ETF geometry, as described before in section 2.4.

However, up to this point, we have not discussed how to model the amplitude and phase shift of the heat wave; which is a function of heater power as shown in Eq. 2.4. Modeling an ETF's amplitude and phase shift allows us to predict its resolution and accuracy, and thus allows us to optimize the readout circuit for area and power. In this section, we attempt to build a general model for ETFs which is also valid for ETFs operating in quasi-ballistic region. Previous models fail to work for small ETFs and generally under-estimate their phase shift and over-estimate their temperature sensitivity [31]. While this model is intended for all ETFs, it was developed and intended for a specific readout architecture (coherent demodulator with square wave drive) in order to simplify the modeling effort.

As a first step, the thermal impedance model developed in [27][33] is used as a baseline. In this model, the thermal impedance between the ETF and the thermopiles is calculated by assuming a semi-spherical distribution of heat in an infinitely large solid. This impedance is calculated for a specific frequency and distance, which is 's' in the case of an ETF. This is a simple and accurate model for ETFs whose hot junctions lie on phase contours, as in [25].

If the ETF heater is driven at a single frequency, we can use equation 2.1 to calculate Z_{ETF} and the temperature distribution on the thermocouples for a given s and α. However, to maximize input power, ETF heaters are usually driven by square-waveforms, which are rich in odd harmonics of the fundamental frequency. Using a square-wave signal to drive the ETF greatly simplifies the required circuitry as described in section 3.4, and is the standard for all published ETFs [5][25][31].

Fortunately, a square-wave can be easily decomposed into harmonics in the frequency-domain. Therefore, a harmonic model was adopted in this work, and the ETF's complex thermal impedance is calculated at both the hot and cold junctions for all relevant harmonics. Thus, for a fundamental frequency of F_{DRIVE}, the following series of impedances are calculated:

$$Z_K = Z_H(KF_{DRIVE}) - Z_C(KF_{DRIVE}) \tag{2.10}$$

Here, K is the harmonic number and the subscripts H and C correspond to the hot and cold junctions respectively. In other to simplify the calculation, only the first 5 odd harmonics of the square wave are taken into account, where K = 1...9. This is justified because an ETF behaves like a low-pass filter.

The calculation of each Z_K impedance, where K is the list of harmonics, is done according to [33], and results in:

$$Z(s,f) = 3\frac{qa * cosh(qa) - sinh(qa)}{4\pi k(T)sa^3q^3} \tag{2.11}$$

Here, $q = \sqrt{j2\pi f/\alpha(s,T)}$, *T* is the temperature, *s* is the distance from heater center, *f* is the harmonic frequency, *a* is the radius of the heater, $\alpha(s,T)$ is the thermal diffusivity, and *k(T)* is the thermal conductivity of silicon. The heater was assumed to be a semi-sphere with a radius of *a*, which is closer to the real implementation for large resistor heaters.

Up to this point, we have ignored the readout structure used for measuring the ETF's phase shift. The choice of readout architecture fundamentally defines how the ETF phase is perceived: as a time-delay that is defined at the zero-crossing of an edge (edge-detection architecture), or as the mean phase of a sine/square wave over one signal cycle (coherent demodulation). Coherent demodulation is the method of choice for high-resolution, low-bandwidth systems as described in detail in section 3.3.2. This method relies on multiplying the ETF waveform with a square-wave demodulation signal at the same frequency and is shown simply in Fig. 2.7.

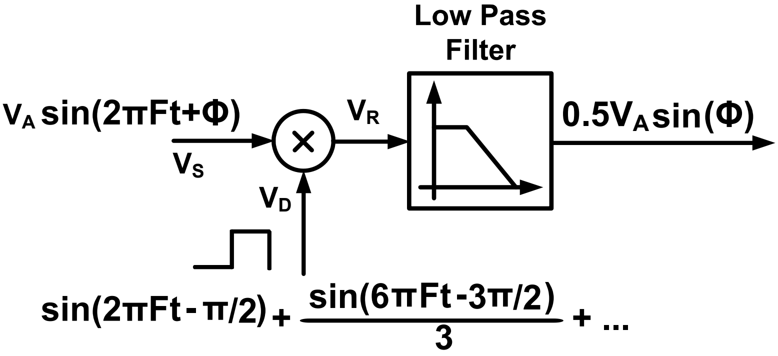


Figure 2.7: Block diagram of a coherent demodulator

In equation 2.10, all harmonic impedances $Z_K, K = 1, 3, 5..$ also contribute to the phase read out at the fundamental, since they are demodulated by a square-wave. After all of the ETF signal's harmonics are multiplied by the demodulating square-wave, the remaining term at DC is a function of the phase shift of the ETF.

We will assume that the ETF heater generates a square-wave heat signal *H(t)* with an amplitude of *A* at a frequency of $\omega/2\pi$. *H(t)* can be expressed in its Taylor series expansion:

$$H(t) = \frac{4A}{\pi}(sin(\omega t + \Phi) + \frac{sin(3\omega t + 3\Phi)}{3} + \frac{sin(5\omega t + 5\Phi)}{5} + ...) \tag{2.12}$$

which can be expressed as,

$$H(t) = \sum_{K=1}^{+\infty} H_K(t) \tag{2.13}$$

Here, $H_K(t) = 0$ when K is even and $H_K(t) = \frac{4Asin(K\omega t + K\Phi)}{K\pi}$ when K is odd. Note that, by expanding *H(t)* to its Taylor series expression, we also observe that the phase of H(t) is equivalent to Φ, which is the phase of its fundamental frequency. Even though the Kth harmonic has K times the phase shift, this is equivalent to the same time delay. Therefore, all of the harmonics of the square wave align in time-domain, and the phase shift of the complete square wave can be defined by Φ alone.

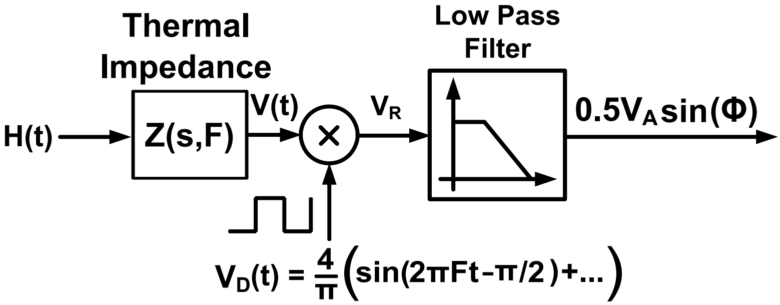


Figure 2.8: Block diagram of coherent demodulator with square-wave drive signal and thermal impedance transfer function Z(f,s)

Fig. 2.8 shows the resulting block diagram with the ETF's thermal impedance and the coherent demodulator. Here, *H(t)* goes through the thermal impedances Z_K, described in frequency domain by $Z(f,s)$ as shown in equation 2.11. Then, the voltage output of the ETF at a specific harmonic K denoted as $V_K(t)$, becomes:

$$V_K(t) = \gamma Z_K H_K(t) \tag{2.14}$$

Here, γ is the Seebeck coefficient. At this stage, the sum of $V_K(t)$ is multiplied by a square-wave signal with $-\pi/2$ phase shift compared to H(t). The phase offset $-\pi/2$ is applied to linearize the non-linearity of the system, further explained in section 3.4. To simplify the result, we'll assume that the filter in Fig. 2.7 is ideal and take a look at the DC term (x) to obtain:

$$x = \gamma \frac{8A}{\pi^2} (\sum_{K=1}^{+\infty} \frac{Z_{AK} sin(K\Phi + \Phi_{ZK})}{K^2}) \tag{2.15}$$

Here, Z_{AK} is the amplitude of Z_K and Φ_{ZK} is the phase shift. Note that the $4/K\pi$ term in equation 2.13 is squared due to the multiplication of two square waves. The phase shift of Z is defined as the value Φ which makes x = 0, as would be the case if purely sinusoidal signals are applied to the coherent demodulator in Fig. 2.7. This can be calculated after Z_K is determined over temperature, drive frequency and other operating conditions. The solved Φ value is then stored as the phase of the ETF at the corresponding condition. This process can be repeated to achieve ETF phase shift over temperature and drive frequency.

The calculation here is an approximation of the coherent demodulation technique described in Chapter 3. More details on this phase readout technique, especially for square-wave drive signals, will be further discussed in section 3.3.2.

For SNR calculations, the RMS of each harmonic is also calculated and added together to obtain the RMS signal strength after demodulation. The RMS signal can be compared to the averaged RMS noise (of the thermopiles, circuit noise etc) to obtain the resolution in phase. Once the master (phase vs temperature) curve of an ETF is calculated, this resolution in phase can be related to temperature error.

We can also take a look at equation 2.15 more intuitively. Here, we can see each harmonic Kth has an approximate weight of the $Z_{AK}/Z_{A1}K$ when compared to the fundamental. This comes from the Z_{AK}/K^2 term and the linear K term inside the sine (when the sine can be linearized for small Φ). Thus, higher harmonics contribute very little to the final phase value. This motivates the choice of including only the first five harmonics (K=1-9) in this model.

From an SNR perspective, higher harmonics contribute even less to signal energy, compared to what they contribute in phase shift. Since the sine function is bounded between -1 and 1, the mean energy of the Kth harmonic has a weight of only $(Z_{AK}/Z_{A1}K^2)^2$ when compared to the fundamental. Thus, only the first and possibly the third harmonic contribute any meaningful signal energy. Ballistic effects can be included into the model by modifying α and Z_K in equation 2.11 as a function of s, as described in section 2.2.1.

Two example ETF designs with s=2 and 3.3 μm will be investigated in section 2.7 to prove the validity of the model. The designs have been implemented in standard 160nm and 40nm CMOS processes, allowing the model to be validated in two different technologies. This allows us to separate the impact of geometrical design of the ETFs from the process and lithographic aspects.

2.6. ETF Accuracy

Before we delve into the details of ETF design, it is important to develop a deeper understanding of thelimits to ETF accuracy. This way, we can predict the inaccuracy of a particular ETF design. The accuracy of an ETF's thermal impedance depends on the precision of the variables in equation 2.11. These are:

1. *s*: Distance from heater center
2. *a*: Heater radius
3. *F*: Drive Frequency
4. α: Thermal diffusivity of silicon
5. *T*: Temperature

The distance *s* is determined by layout, and, as a lateral dimension, its precision is determined by lithography. Generally, the effect of *a* is small, so its accuracy can be neglected. For a typical compact ETF design with s > 2 μm and a < 0.5 μm, the effect of *a* is roughly an order of magnitude smaller than *s*. The precision of *F* can be improved by choosing a good frequency source, which is used as the timing reference for ETF measurements.

The thermal diffusivity of IC grade silicon is quite well defined. This is due to the fact that α is determined by the fundamental mechanical properties of a silicon crystal. Previous studies on the effect of doping on α [5] show that α is not sensitive to low-level doping, such as in an n-well. However, since α is a mechanical property, it may be susceptible to mechanical stress. Measurements done in [32] demonstrate that α is accurate at least to an equivalent temperature of 0.2 ^{o}C (3σ) for low-stress (ceramic packaged) devices.

An ETF consumes significant amount of power and increases its temperature (T) compared to the ambient. This self-heating of an ETF will vary according to the spread of the heater power; and thus introduces an error.

In the following sections, three inaccuracy sources will be discussed: lithography, mechanical stress and self-heating. Lithography is a limit for older technologies, while self-heating becomes an issue for smaller ETFs. Mechanical stress is more important for volume production, when ICs undergo significant stress after plastic packaging.

2.6.1. Lithography

Errors in lithography will cause the critical distance *s* of an ETF to spread, leading to a phase shift in the complex thermal impedance Z_K in equation 2.11. As a result, the phase shift of an ETF (Φ) will also spread over temperature.

The spread of *s* can be estimated from the inaccuracy of the lithography or more specifically, the accuracy of the process masks used to build the ETF. The accuracy of lithography is expressed via several parameters, such as critical dimension (CD) uniformity and overlay [34]. CD uniformity (3σ) error is the deviation of a single-mask, single-line width feature from its mean value, while overlay error is the absolute positional error between a mask feature and its placement on the wafer. Note that overlay errors include alignment errors between different masks and are thus much larger than CD uniformity error. An ETF's phase shift is mostly defined by a relative dimension quantity *s*, which is ideally the distance between structures defined by the same mask (rather than alignment of two masks), CD uniformity error is more relevant to ETF design.

Fig. 2.9, from the ITRS 2007 roadmap [34] shows the recommended CD uniformity error and overlay error bounds for sub-65nm processes. The CD uniformity error is bounded between 7% to 12% of the physical gate length for most processes, and is closer to 12% for most modern logic and mixed-signal processes [35]. In order to reduce costs, CD uniformity error are reported and heavily monitored for the gate polysilicon layer, but not so strictly for other layers [35].

Moreover, the ETF must be carefully laid out to avoid two separate masks from defining the critical ETF geometry *s*. A typical BAR ETF layout is shown in Fig. 2.10, where two masks define the ETF geometry: the active (p+) and silicide protection masks. Active regions that are covered with the silicide protection mask are p+ doped resistors, while regions around contacts are coated with silicide. These regions exhibit very low resistivity and Seebeck coefficient [36]. Therefore, the entirety of these regions act as the hot (and cold) junctions of the thermopiles, rather than the smaller contact mask. In other words the cold/hot junctions exist at the interface between the silicide and the p+ doped regions, and not between the metal contacts and the silicide. The distances *S0* and *S1* in the figure show the closest and farthest end of the hot junction from the mid-point of the heater. Thus, we can approximate *s* as (*S0*+*S1*)/2.

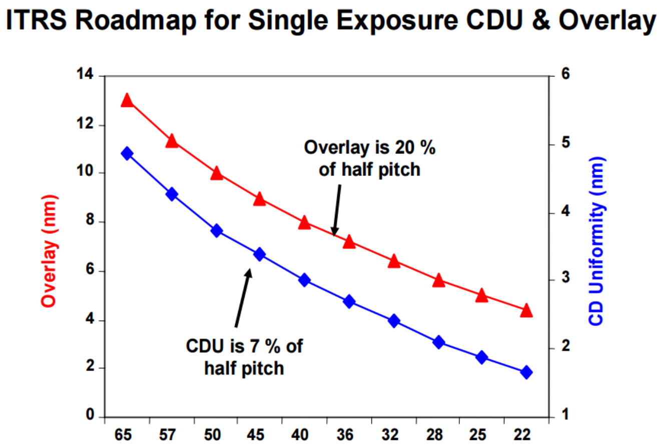


Figure 2.9: Recommended CD uniformity and overlay error bounds for sub-65nm processes

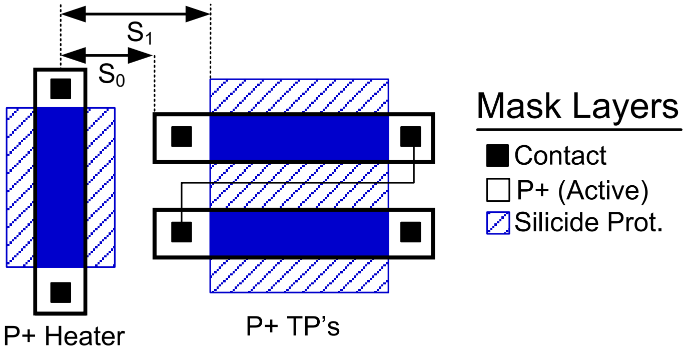


Figure 2.10: Three masks showing the composition of a simple BAR ETF. Dark blue regions are p+ resistors, dark black squares are contacts and un-shaded rectangular regions are silicided p+ active regions.

The distance *S0* is determined only by the P+ active mask, while *S1* is defined by both the P+ active and the silicide protection masks. For this ETF, both overlay and CD uniformity error on both masks will contribute to its inaccuracy.

It is possible to lay out the ETF such that only the silicide protection layer determines *S0* and *S1*, as shown in Fig. 2.11. This is achieved by extending the silicide protection layer mask of the heater resistor. Note that, provided the thermocouples are symmetrically distributed around the heater, positional errors of the heater itself will be averaged out. This is due to the fact that, as the heater gets further away from one hot junction, it moves closer to the junction on the opposite side of the phase contour.

Hence, only the CD uniformity error of the silicide protection mask contributes to inaccuracy. Unfortunately, this error is not directly related to the gate CD uniformity error, which will typically be more tightly controlled.

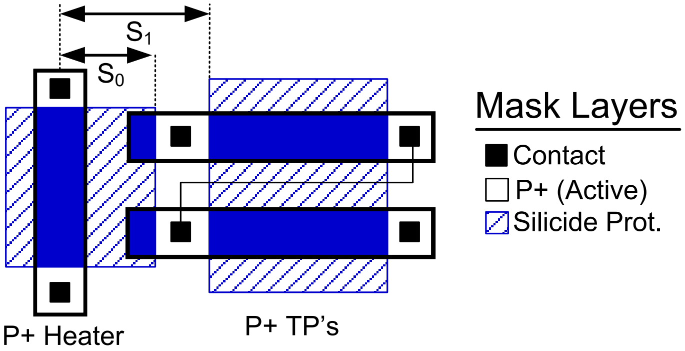


Figure 2.11: Three masks showing the composition of a simple BAR ETF, where only the silicide protection mask defines the critical distance *s*

To predict the inaccuracy due to lithography errors in a particular process, we will assume that the best possible mask tool for the silicide protection layer is used. This tool would also define the minimum pitch of the first metal layer, which is heavily optimized to reduce the area of digital standard cells. In reality, a less advanced mask tool may be used for the silicide protection mask, since it is not a critical layer for commonly used digital blocks. With this in mind, we can approximate Δs, or the 3σ error on *s* due to mask CD uniformity error as:

$$\Delta s \approx 3\sigma_{CDU} * HP_{M1} \tag{2.16}$$

Here, σ_{CDU} is the normalized standard deviation of CD uniformity, and HP_{M1} is the half-pitch of metal 1 layer in the process, also known as half of the SRAM pitch. As described before, σ_{CDU} ranges between 2.33% to 4% for a typical CMOS process. For a 40nm CMOS process with a metal 1 half-pitch of 63nm and assumed σ_{CDU} of 4%, Δs is found to be 7.56 nm (3σ). Using the methodology described in section 2.5, we can calculate the change in phase shift $\Delta\Phi$ due to an error of Δs in the ETF. From the ETF's master curve we can then convert the phase error into temperature inaccuracy. For an s = 3.3 μm ETF (denoted as an s3.3 ETF) driven at 1.17 MHz (see Fig. 2.12), Δs = 7.56 nm is equivalent to an inaccuracy of 0.7 °C (3σ). Here, Δs is estimated as the global variation of s, rather than the average variation of individual hot-junction locations. Δs = 7.56 nm is close to the experimental results obtained (0.85 °C, 3σ) with a test chip (see Chapter 5) in a standard 40nm process, once room-temperature trimming was applied to remove the remainder of readout-related errors.

The estimation in equation 2.16 will be used to predict the spread of ETFs presented in sections 2.7.1 and 2.7.2. We will consider a technology with metal1 half-pitches of 63nm and σ_{CDU} of 4%, roughly corresponding to a standard 40nm CMOS processes.

Even though the gate length continues to scale in modern CMOS, the 2011 ITRS roadmap notes that the half-pitch dimension for single-exposure masks is practically limited to 40nm [35]. It can be safely predicted that the silicide protection layer will not be migrated to multi-patterned masks to save costs, and hence future scaled ETFs may hit a wall with respect to accuracy. This boundary can be calculated via equation 2.16, and it corresponds to roughly 0.45 °C (3σ) for an s = 3.3 μ m ETF. ETF dimensions will need to be defined with only the critical mask layers (gate poly, metal, active) to circumvent this problem.

Another important question concerns the variation of Δs over large areas of a single wafer, or between multiple wafers. Since CD uniformity error calculated by the foundries can cover errors over multiple lots, it is difficult to differentiate between wafer-to-wafer and inter-wafer error sources. In the future, more batch-to-batch measurements will need to be done to further understand this problem.

2.6.2. Self Heating

An ETF's heater not only generates a dynamic heat signal at F_{DRIVE}, but also a DC heat signal that increases die temperature. The self-heating error at the thermopile hot-junction can be calculated from equation 2.5, together with the thermal impedance of the package Z_{PKG}. This results in a DC thermal impedance of:

$$Z_{DC} = \frac{T_{DC}}{H_{DC}} = \frac{1}{2\pi ks} + Z_{PKG} \tag{2.17}$$

For a given heater power H_{DC} and Z_{PKG}, the self-heating T_{DC} can be calculated from equation 2.17. A typical plastic package (such as SO28) will exhibit a thermal impedance of $\sim$ 100 °C/W. With H_{DC}=2.5mW and s = 2 μm, Z_{DC} = 630 °C/W, and T_{DC} is 1.5 °C. For a typical ceramic DIL28 package with Z_{PKG} = 11 °C/W, Z_{DC} = 581 °C/W and T_{DC} is 1.3 °C. The variation of Z_{PKG} may be a significant source of error, as it depends on the PCB design and mechanical tolerances.

While we can calculate the self-heating at the hot junction in this manner, the ETF's thermal diffusivity (α) will also be influenced by the average self-heating of the silicon in-between the heater and the hot junction. To first-order, this is given by the following integral:

$$Z_{TD} = \frac{1}{s}\int_0^s \frac{1}{2\pi ks}ds + Z_{PKG} \tag{2.18}$$

Note that taking the integral of equation 2.18 with boundaries 0 and s results in infinite Z_{TD} arising from the $ln(0)$ term. This goes back to the previous discussion about ballistic phonon transport, and how the heat diffusion theory breaks down for small distances in silicon. Taking this into account, a better approximation is given by:

$$Z_{TDbal} = \frac{1}{s - smin}\int_{smin}^s \frac{1}{2\pi k(s)s}ds + Z_{PKG} \tag{2.19}$$

Where $smin$ = 0.5um is the distance over which most of the heat transport is ballistic, and $k(s)$ is the thermal conductivity of silicon as a function of heat transport distance. Assuming that the thermal capacitance C is constant, Fig. 2.2

can be used to determine the relationship of k with s as well. Setting a minimum boundary for $smin$ allows us to disregard the impact of ballistic transport on Z_{TDbal}, since the two heat transport mechanisms are unrelated [9]. Regardless, the term $k(s)$ in equation 2.19 makes the calculation complicated, and its precise derivation goes back to the experiments done in [9] and [12]. To simplify the relationship, we will equate $k(s)$ to $k * c$, where c is the mean of $k(0.5um)$ and $k(s)$ as calculated from figure 2.2 and [9] [12].

For an s3.3 ETF and $smin$ = 0.5μm, c = 0.525 and ignoring the impact of Z_{PKG}, Z_{TDbal} is numerically calculated as:

$$Z_{DCs33} = \frac{1}{2\pi ks} * \frac{ln(3.3) - ln(0.5)}{(3.3 - 0.5) * 0.525} = 1.2837 * Z_{DC}(3.3) \tag{2.20}$$

Where $Z_{DC}(3.3)$ is the result of equation 2.17 for a hot junction distance of 3.3μm. Therefore, the impact of self-heating for a compact ETF can be estimated from Z_{DC} combined with a gain factor, defined here as g. As the distance decreases, this gain factor is expected to increase dramatically. An s2 ETF for example, has c = 0.41 and g = 2.1; which means the self-heating of the heat diffusion path is estimated to be twice that of the self-heating of the hot junction. This means that self-heating could be a major s for small ETF designs.

Even if Z_{DC} was a process-independent parameter, similar to an ETF's Z, H_{DC} can vary significantly between different chips. If the ETF heater is supplied by a constant voltage source, then its resistance dominates this spread. This is typically about ±20%. With this rule of thumb, we can estimate from equation 2.20 that the spread due to self-heating is about 0.33 °C.

Self-heating is a bigger problem for high-resolution ETFs, or when power consumption is increased. It can be solved by precisely regulating the ETF's power consumption, but this brings additional complexity to the system.

2.6.3. Mechanical Stress

BJT based circuits, such as temperature sensors, are sensitive to mechanical stress [37]. This is especially problematic for devices that have been trimmed before packaging, since most plastic packages cause considerable stress. During encapsulation, hot glue or epoxy is used to cover the chip; and after packaging this epoxy is left to cool down. The mechanical properties of the epoxy are different before and after cooling, and hence the epoxy applies stress to the encapsulated chip. The exact amount of stress introduced by packaging can vary a lot depending on the material and package type.

Thermal diffusivity based temperature sensors have been thought to be immune to the effects of mechanical stress; however new research and experimental results (see Chapter 5) seem to indicate that they are also affected by mechanical stress.

Stress directly influences the thermal diffusivity of silicon by altering the distance between the atoms in its crystal lattice, thereby altering both k and α. In fact, k can vary by a factor of 2 over a ±5% compressive and tensile strain [38]. We will assume that α changes similarly. Silicon is an elastic material for small compressive or tensile strains, and hence we can relate strain and stress linearly via:

$$\epsilon = \frac{\sigma}{E} \tag{2.21}$$

Here, ϵ is the unit-less measure in strain, σ is stress in pascals (Pa) and E is the material's Young's modulus, also in units of Pa. For silicon, E depends on its crystal orientation [39]. We will assume it is 150 GPa for a <100> orientation wafer (most common wafer type for IC processes), although it can vary by ±30% according to the orientation of the stress vector with respect to the wafer surface. Referring to equation 2.21 and [38], we can calculate that 100 MPa of stress changes k and α by 0.67%. For an s = 3.3 μm ETF, this results in 2 °C error at room temperature. If the strain is compressive, α increases and hence the ETF under-estimates ambient temperature; while if the strain is tensile, then the ETF will over-estimate.

The error in α due to stress is a percentage or gain error, which means the error correlates directly with temperature. Hence, a proportional-to-temperature (PTAT) or gain trim could be used to eliminate stress-related inaccuracy.

2.7. Polygon ETF Designs and Performance in Standard 160nm and 40nm CMOS

In this section, the performance of two small polygon ETFs will be presented. It will be shown that, combining the harmonic impedance model presented in section 2.5 with quasi-ballistic transport effects results in a model that agrees well with measurement results. The objective of this section is to provide a thorough analysis and understanding of these ETFs, as a first step in the design of a TD sensor, by calculating their optimum drive frequency, RMS signal level and thermal resolution.

2.7.1. A scaled s=3.3 μm ETF

The ETF presented in [31], with s=4.7 μm is a good starting point for a high-SNR, lower accuracy ETF suitable for thermal management applications. In this work, however, an s=3.3 μm ETF was chosen to increase signal strength by x2, at the cost of only 40% worse accuracy compared to s=4.7 μm. The x2 difference in signal strength comes from simulations done with the aforementioned harmonic thermal impedance model. In order to maximize signal strength, the number of thermopiles, or N, was set to 16. This was limited by the DRC restrictions of the 160nm CMOS process. From this point onward, this design will be referred to as an s3.3 ETF.

Two versions of this ETF have been taped out in two different processes. The two versions involve the placement of the cold junction (CJ) at 7.1 μm and 11 μm respectively. Since their cold junctions are further away, the ETFs with CJ=11 μm achieve better SNR, but occupy 2.5x more area (600 μm^2 compared to 240 μm^2). Hence, when the ETF designs were migrated from 160nm to 40nm CMOS, only the smaller ETFs with CJ=7.1 μm were realized. The two ETFs will be referred to as CJ7.1 and CJ11 ETFs in the rest of this section.

The thermal impedance model used to predict the behavior of this ETF assumes a point heater element, which is not practical. While previously, compact ETFs have used U-shaped diffusion(n+ or p+) heaters [31], the s3.3 ETF uses a compact bar-shaped heater. This is facilitated by the low resistivity of diffusion resistors, which is in the range of 100-200 Ohms/square. A heater made from a single square will then dissipate a few milliwatts of power from a 1V supply, making it quite suitable for use in a ETF. Depending on the process, the heater will then have minimum dimensions in the range of 0.5-um per side, making it a good approximation to a point source in an s3.3 ETF.

The simulated phase vs. temperature behavior (also known as a master curve) of the two ETFs is shown in Fig. 2.12 and 2.13, respectively. Simulations were done over 4 drive frequencies, at harmonics of 585.9 kHz to demonstrate how the master curve changes with respect to drive frequency. 585.9 kHz was chosen as the baseline drive frequency because it can be easily obtained from a commonly available 75 MHz clock source by division. For the simulations, the harmonic thermal impedance model with the ballistic effects was used. Quasi-ballistic transport was modeled by using Fig. 2.2 to modify α for $s = 3.3\mu$m hot junction and the cold junctions at 7.1 and 11 μm. Assuming that for bulk silicon α = 0.88 cm$^2/s$, this results in 0.66 cm$^2/s$ for the hot junction, 0.79 cm$^2/s$ for the 7.1 μm cold junction and 0.84 cm$^2/s$ for the 11 μm cold junction. Since the α at cold junctions is close to the bulk value, the impact of ballistic transport was found to be negligible for them and the bulk silicon value of 0.88 cm$^2/s$ was used for calculating thermal impedance at the cold junctions. The hot junction uses α = 0.66 cm$^2/s$ to account for quasi-ballistic transport.

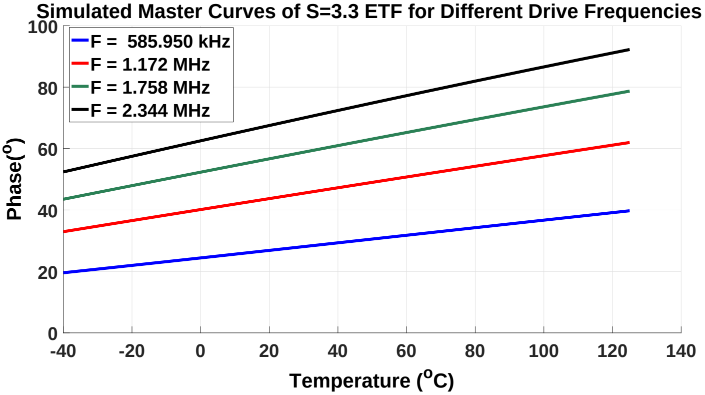


Figure 2.12: Phase vs. Temperature (Master) Curve of an s3.3,CJ7.1 ETF at different drive frequencies

2.7.1.1. RMS Signal and Resolution of s=3.3 μm ETF

We can also obtain the RMS signal of the ETFs from the model. Figures 2.14 and 2.15 show the RMS signal strength of CJ7.1 and CJ11 ETFs respectively. The Seebeck coefficient was estimated from measurements as 0.25 mV/K, and as such this value was used in the model. The mean heater power was set to 2.5 mW, where the heater signal is a 0 to 5 mW square wave with 50% duty cycle, driven at

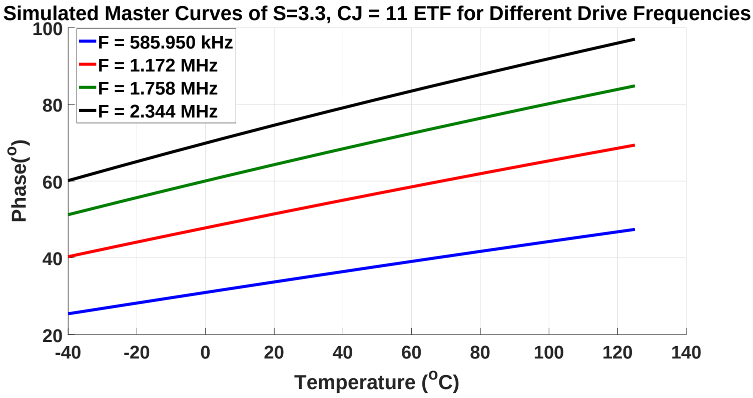


Figure 2.13: Phase vs. Temperature (Master) Curve of an s3.3, CJ11 ETF at different drive frequencies

the drive frequency. To dissipate this power, the heater resistance was set to 550 Ω in a 160-nm 1.8V process and 190 Ω in a 40nm 1.2V process. Due to parasitic resistance (from contact regions and interconnects) and switch losses, the final power delivered to the ETF is always less than that calculated using the heater resistance alone. Hence the heater voltage and power dissipation must be adapted during measurement to evaluate the ETF's SNR accurately.

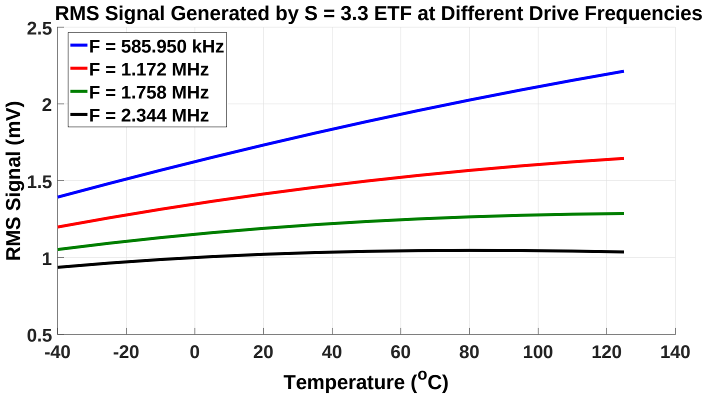


Figure 2.14: RMS Signal Strength of an s3.3/CJ7.1 ETF over drive frequency and temperature, for 2.5mW heater power

Shifting the distance between hot and cold junctions from 7.1 to 11 μm increases the RMS signal by 20%, and the thermal noise by 16% due to larger geometry. Thus, it may seem that both ETFs have similar SNR, however the CJ7.1 ETF is more susceptible to circuit noise due to its smaller signal strength. This susceptibility depends on the thermopile resistivity and circuit noise. For a thermopile resistivity of 150 Ω per square, we get a resistance of 6 kΩ for the s3.3/CJ7.1 ETF and 8 kΩ for the s3.3/CJ11 ETF. Assuming a circuit noise equivalent to 1 kΩ, we get a noise

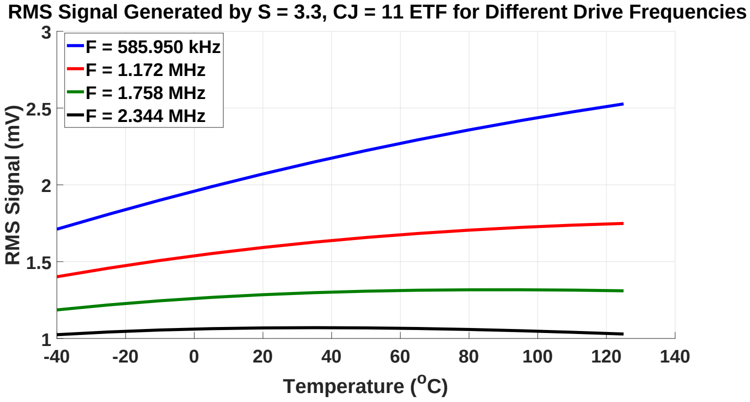


Figure 2.15: RMS Signal Strength of an s3.3/CJ11 ETF over drive frequency and temperature, for 2.5mW heater power

penalty of 13% and a net improvement in SNR of 7% for the s3.3/CJ11 ETF when compared to s3.3/CJ7.1.

Also, note that the RMS signal strength drops significantly as frequency increases, and that the RMS signal is not a monotonically increasing function of temperature for all frequencies. The latter is because the thermal conductivity (k) in equation 2.5 is also a function of temperature. This is taken into account via the approximate relation $k \propto T^{-1.3}$ [6], where T is the absolute temperature. This means that, for particular k, s, T and F values; any change in the exponential term in equation 2.5 may be partially compensated by the linear term. This means that the slope of the lines in Figures 2.14 and 2.15 do not have to be positive, and can vary as a function of T and F.

Figure 2.16 shows the estimated resolution of an s3.3, CJ11 ETF for a heater power of 2.5 mW, conversion time of 1ms, and a readout with g_m=1mA/V. The ETF's thermopile resistance is 8 kΩ as reported before, and hence its thermal noise dominates over the readout. These results are swept over frequency and temperature to choose the optimum drive frequency, and it is shown that the range of frequencies between F=1.172-1.758 MHz achieves optimum resolution. This is despite the higher signal amplitude at lower frequencies (such as 585.950 kHz), since at such frequencies the ETF phase shift is much less sensitive to temperature, as was shown in 2.13.

In the end, F=1.172 MHz was chosen as the frequency of choice for both s3.3 ETFs; since using a frequency on the lower end of the optimum resolution range also relaxes the bandwidth requirement of the readout. Note that the associated resolution of 0.24 ^{o}C (RMS) of s3.3 matches the experimental results (see Chapter 4) for F=1.172 MHz and g_m=1mA/V. This validates the usefulness of the model in predicting ETF resolution.

2.7.1.2. Quasi-Ballistic Transport and s=3.3 μm ETF

At this point, we will take a look at how important the ballistic effects are for

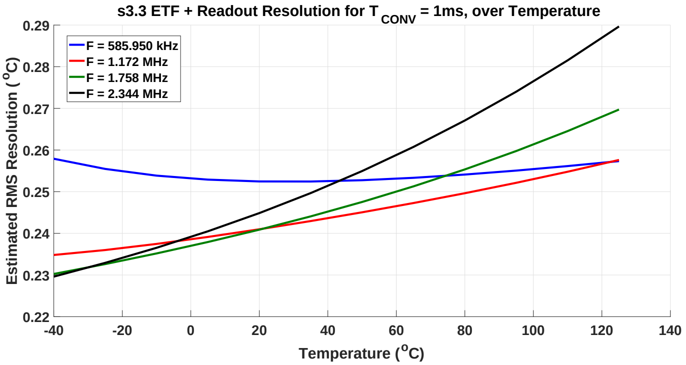


Figure 2.16: Estimated resolution of an s3.3 ETF for 2.5 mW heater power, conversion time of 1ms, and readout g_m=1mA/V

the accuracy of the model, and how relevant they are to experimental results. Therefore, we compare silicon data (see Chapter 5) with the model results, with and without the inclusion of ballistic effects. This can be done by either making α constant or re-defining it as a function of s; as was shown in section 2.2.1. Figure 2.17 shows a comparison of s3.3 ETF's measured and simulated master curves at F=1.172 MHz. The model seems to be in good agreement with measurements, but it over-estimates the ETF's temperature sensitivity by 10% above 60^{o}C, and under-estimates phase shift by 2 ^{o}C below. This can be due to various un-modeled factors, such as readout errors, consistent mechanical stress at hot temperatures, thermal contribution of STI, oxide and epi layers and variation of ballistic effects as the temperature increases.

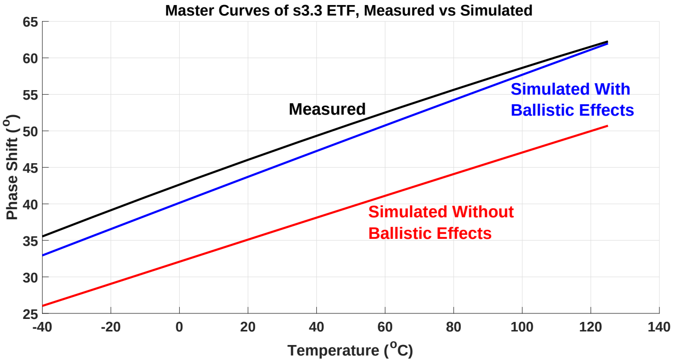


Figure 2.17: Comparison of measured s3.3 ETF master curve, with simulated curves with and without ballistic effects

Nevertheless, if the ballistic effects are turned off, the model under-estimates ETF phase shift by roughly 9^{o}C. As shown in Fig. 2.18, it also over-estimates RMS signal amplitude by 15%. Both of these errors are large enough to warrant the inclusion of ballistic effects. Note that there is no experimental data for Fig. 2.18. This is because with the current implementation of the phase-domain readout architecture, it is not possible to determine the actual amplitude or RMS strength of the ETF signal.

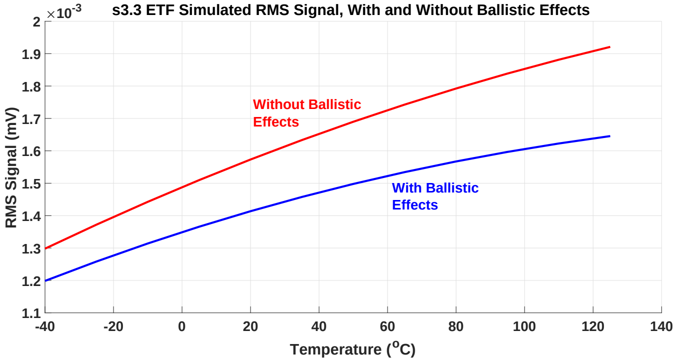


Figure 2.18: Comparison of s3.3 ETF RMS signal amplitude, with and without ballistic effects

2.7.1.3. Inaccuracy of s=3.3 μm ETF

Finally, we can calculate the estimated inaccuracy of this ETF, via the lithography spread model presented in section 2.6.1, equation 2.16. Figure 2.19 shows the estimated 3σ spread of an s3.3 ETF for different frequencies and for a half-pitch of 63nm. Variation of s causes a proportional-to-temperature (PTAT) error that mirrors the temperature behavior of an ETF. The plot can be easily scaled to different CMOS technologies with a known metal half-pitch and σ_{CDU}. For example, in a more mature 160nm CMOS process with a metal half-pitch of 256nm, the worst-case inaccuracy can be expected to be within a range of values from 1.8 to 3 °C, depending on σ_{CDU} for the silicide protection mask. The mean estimate of 2.4 °C is identical to the measured 2.4 °C (3σ, untrimmed) inaccuracy from 96 samples [40].

For a 40nm process with a half-pitch of 63nm (see Chapter 5), the measured inaccuracy was 1.4 °C for (3σ, untrimmed); which under-estimates the model by a factor 2 (see Fig. 2.19). The inaccuracy drops to 0.75 °C after one-point trimming, which fits the model better. This difference in measured and modelled accuracy might mean that either the readout dominates the phase error (and the ETF error is only visible after one-point trim) or that the half-pitch of 63nm is too optimistic for an ETF in 40nm process. To obtain 1.4 °C inaccuracy due to lithographic spread only, the half-pitch would have to be 118nm.

2.7.2. A scaled s=2 μm ETF

The improved lithography in modern processes (such as 40nm CMOS) can be leveraged to design even smaller ETFs, which trade-off favorably in resolution over accuracy. To investigate this trade-off, an ETF with s=2 μm was designed with 2x

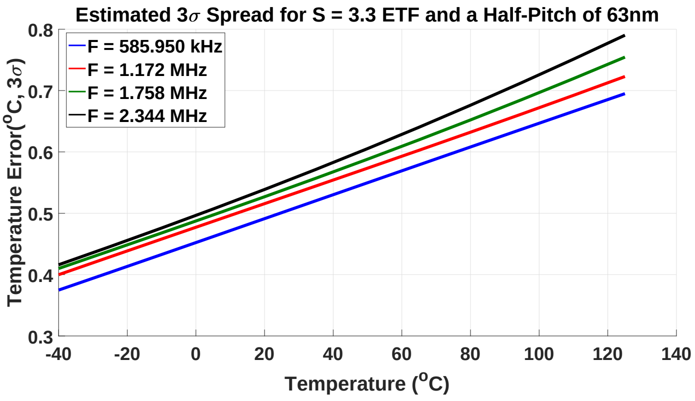


Figure 2.19: Estimated spread of an S3.3 ETF for a half pitch of 63nm

higher signal strength and 63% better SNR, at the cost of 65% worse accuracy. For compatibility reasons, the same cold junction distance of 7.1 μm and N = 16 parameters of the s3.3 ETF was adopted. Heaters were implemented with minimum size square-shaped diffusion resistors, similar to s3.3 ETF. We will refer to this ETF as an s2 ETF. The master curves of the s2 ETF at different drive frequencies are shown in Fig. 2.20. Similar to the s=3.3μm ETF, all simulations were done with the harmonic thermal impedance model including quasi-ballistic effects, unless otherwise specified.

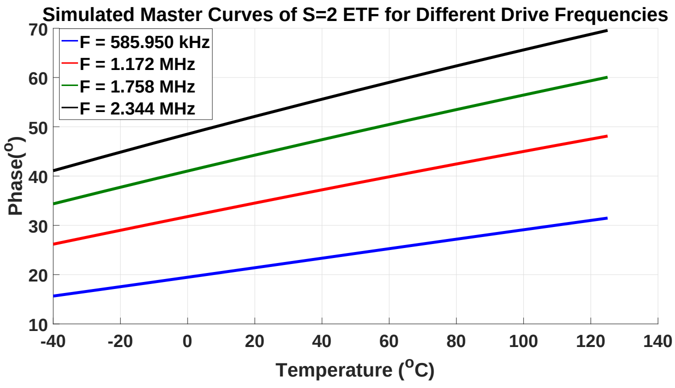


Figure 2.20: Phase vs. Temperarture (Master) Curve of an s2 ETF at different drive frequencies

Figure 2.21 shows the RMS signal strength of an s2 ETF, for a Seebeck coefficient of 0.25 mV/K and heater power of 2.5 mW. When compared to an s3.3 ETF, we can see that the signal strength is roughly doubled. However, due to longer thermopile arms and addition of a highly-dense resistive area near the heaters, thermopile resistance and noise increases by 50% and 22% respectively. In a standard 40nm CMOS process with thermopile resistivity of 215 Ω per square, this results in a 12 kΩ thermopile resistance. Similar to the s3.3 ETF, a heater resistance of 190 Ω was chosen for 1.2V operation.

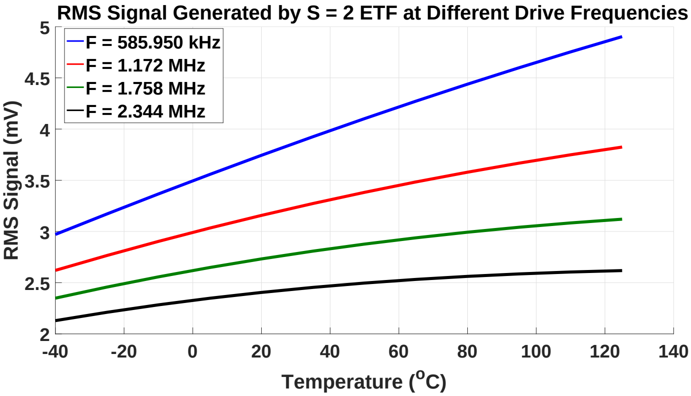


Figure 2.21: RMS Signal Strength of an s2 ETF over drive frequency and temperature , for 2.5mW heater power

Figure 2.22 shows a comparison of the simulated and measured master curves of the s2 ETF. Similar to the s3.3 case, the model over-estimates the ETF's temperature sensitivity by 10% but roughly captures the ETF's phase over temperature. The gap between measured and simulated master curves without ballistic effects is even larger for the s2 ETF, and is close to 11 ^{o}C. This is expected: in reality α reduces with s, and hence standard models with constant α will under-estimate ETF phase shift more.

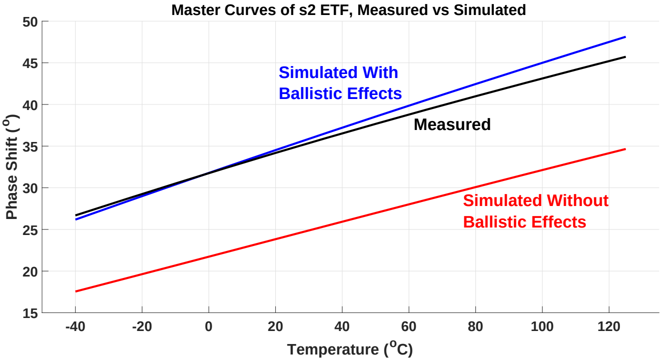


Figure 2.22: Comparison of measured s2 ETF master curve, with simulated curves with and without ballistic effects

In terms of accuracy, an s2 ETF exhibits 65% worse spread when compared to s3.3 ETF, as shown in Fig. 2.23. Here, the same methodology as for the s3.3 ETF via equation 2.16 was adopted. For implementation in a 40nm CMOS process, a half-pitch of 63nm was assumed for the model. Due to its small size and DRC restrictions, such an ETF can only be built in advanced (65nm and below) technologies, and hence benefits from process scaling to offset its higher inaccuracy. The predicted spread of 1.2-1.4 °C underestimates the measured untrimmed inaccuracy of ±2.3°C (see Chapter 5), but is close to the trimmed inaccuracy of ±1.05 °C.

Using previous comparison of the s3.3 ETF model and measurement data, we can calculate what the inaccuracy estimate is for a half-pitch of 118nm, which was obtained via fitting the measurement data into the lithographic inaccuracy model. For an s2 ETF, this corresponds to 2.4 °C and matches very well with the measurement data. This suggests that the half-pitch of 40nm process may not be 63nm as expected, but is roughly double for a polygon ETF structure.

Further work including extra experimental data, over more ETFs in different processes would be necessary to exhaustively evaluate if this is indeed the case and the benefit of technology scaling for ETFs is worse than expected. The alternative explanation for the difference between the modeled and measured inaccuracies could be due to additional parasitic capacitance introduced by the polygonal ETF geometry; which adds significant electrical phase shift. This is discussed further in the next section.

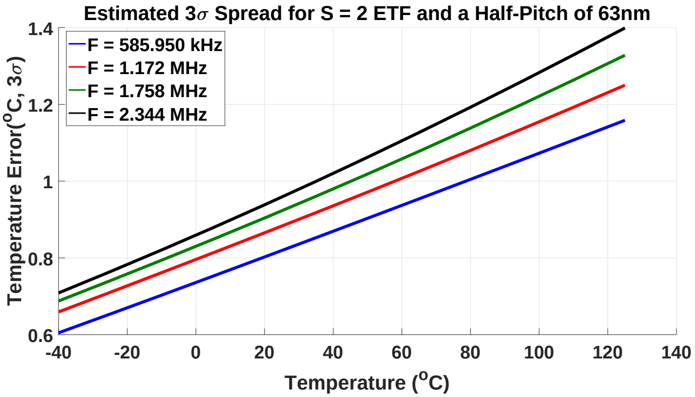


Figure 2.23: Estimated spread of an S2 ETF for a half pitch of 63nm

2.8. Summary of ETF Design Trade-offs

To summarize this chapter on ETF design, we can condense all design effort into trade-offs between the ETF's energy efficiency (resolution/power) and its accuracy. This manifests itself in various stages of the design, starting with the choice of ETF geometry. As explained in section 2.4, choosing a polygon geometry results in the best resolution and energy efficiency, but greatly increases the ETF's electrical phase shift due to the large parasitic junction capacitance between thermopiles and substrate.

The second trade-off between accuracy and resolution, other than the choice of phase-contour vs polygonal ETF geometries, arises from s. Here, s must be chosen low for best energy efficiency, but a small s can ultimately define the ETF's accuracy due to limited lithography resolution. Since this is a hard limit on accuracy, s must be chosen according to the lithography resolution and the sensor's accuracy requirements. If the sensor is to be calibrated over temperature, then these requirements can be relaxed. However, there are no theoretical guidelines on how much temperature calibration can improve the ETF's accuracy. If the results presented in Chapter 5 are used as a guideline, an offset trim done at room temperature improves accuracy by 2x.

The third important design variable is the drive frequency. When the drive frequency increases, the readout's finite bandwidth introduces a larger delay and hence the sensor's accuracy degrades as well. Therefore, choosing a lower frequency is always best for accuracy purposes. On the other hand, as shown in section 2.7.1, there exists an optimum drive frequency (or a range of frequencies) for achieving the best energy efficiency. We choose the drive frequency based on energy efficiency concerns, since the readout circuit's bandwidth can be improved by consuming more power. This is not a problem since the ETF typically burns 5-10x more power than its readout [25][31][32].

2.9. Conclusions

The analysis, simulation and design of scaled ETFs have been described in this chapter. For high SNR applications, a novel polygon ETF that reduces thermopile resistance by 1.5X is described. A method that models the ETF's complex thermal impedance over multiple drive harmonics was used to design two ETFs with s = 2 and 3.3 μm. Ballistic transport effects, which increase the mean phase shift of ETFs over temperature have been included in the model. With these efforts, it was possible to improve the scaled ETF's SNR and area for fast and compact temperature sensors that are useful in thermal management applications. When compared to the experimental results in Chapters 4 and 5, the simulation results have been found to correctly predict the accuracy, SNR and phase-vs-temperature behavior of ETFs in silicon.

The second part of the chapter discusses inaccuracy sources for ETFs. Errors due to limited lithography resolution during manufacturing, self-heating during operation, and mechanical stress related errors are discussed. Lithographic error sources dominate an ETF's accuracy, especially if a gain or PTAT trim is used to remove the effects of mechanical stress. For future work with even smaller ETFs, regulating the ETF's self-heating will be needed to achieve good accuracy.

References

[1] M. A. P. Pertijs, K. A. A. Makinwa, and J. H. Huijsing, "A CMOS Smart Temperature Sensor With a 3σ Inaccuracy of ± 0.1 ^{o}c from -55 ^{o}c to 125 ^{o}c," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 12, pp. 2805–2815, Dec 2005.

[2] B. Yousefzadeh, S. H. Shalmany, and K. Makinwa, "A BJT-based temperature-to-digital converter with (3σ) inaccuracy from -70 ^{o}C to 125 ^{o}C in 160nm CMOS," in *2016 IEEE Symposium on VLSI Circuits*, June 2016, pp. 1–2.

[3] M. G. Holland, “Analysis of Lattice Thermal Conductivity,” *Phys. Rev.*, vol. 132, pp. 2461–2471, Dec 1963.

[4] H. R. Shanks, P. D. Maycock, P. H. Sidles, and D. G. C., “Thermal Conductivity of Silicon from 300 to 1400 ^{o}K,” *Phys. Rev.*, vol. 130, pp. 1743–1748, Jun 1963.

[5] C. van Vroonhoven and K. Makinwa, “Thermal Diffusivity Sensors for Wide-Range Temperature Sensing,” in *2008 IEEE Sensors*, Oct 2008, pp. 764–767.

[6] H. R. Shanks, P. D. Maycock, P. H. Sidles, and G. C. Danielson, “Thermal Conductivity of Silicon from 300 to 1400°k,” *Phys. Rev. Letters*, vol. 130, pp. 1743–1748, Jun 1963.

[7] Thompson J. C. and Younglove B. A., “Thermal conductivity of silicon at low temperatures,” *J. of Phys. Chem. Solids*, vol. 20, pp. 146–149, 1961.

[8] D. Hilbiber, “A new semiconductor voltage standard,” in *IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, vol. VII, Feb 1964, pp. 32–33.

[9] K.T. Regner, et. al., “Broadband phonon mean free path contributions to thermal conductivity measured using frequency domain thermoreflectance,” *Nature Communications*, vol. 4, no. 1640, 2013.

[10] P.G. Sverdrup, S. Sinha, M. Asheghi, S. Uma, and K.E. Goodson, “Measurement of ballistic phonon conduction near hotspots in silicon,” *Applied Physics Letters*, vol. 78, no. 21, pp. 3331–3333, 2001.

[11] G. Chen, “Thermal conductivity and ballistic-phonon transport in the cross-plane direction of superlattices,” *Phys. Rev. B*, vol. 57, pp. 14 958–14 973, Jun 1998.

[12] J.A. Johnson et. al., “Direct Measurement of Room-Temperature Nondiffusive Thermal Transport Over Micron Distances in a Silicon Membrane,” *Phys. Rev. Lett.*, vol. 110, p. 025901, Jan 2013.

[13] G. Chenand and C. Dames, *Thermal Conductivity of Nanostructured Thermoelectric Materials*. CRC Press, 2005, pp. 42–1–42–16.

[14] A. J. Minnich, “Determining Phonon Mean Free Paths from Observations of Quasiballistic Thermal Transport,” *Phys. Rev. Letters*, vol. 109, p. 205901, Nov 2012.

[15] R. B. Wilson and C. D. G., “Anisotropic failure of Fourier theory in time-domain thermoreflectance experiments,” *Nature Communications*, vol. 5, Oct 2014.

[16] H. J. Eichler and A. Hermerschmidt, *Photorefractive Materials and Their Applications I*. Springer, 2006.

[17] J. Ebrahimi, "Thermal diffusivity measurement of small silicon chips," *Journal of Physics D Applied Physics*, vol. 3, pp. 236–239, Feb. 1970.

[18] Y. C. K. Souri and K. A. A. Makinwa, "A CMOS Temperature Sensor With a Voltage-Calibrated Inaccuracy of $\pm$ 0.15 ^{o} c (3σ) from -55 ^{o}c to 125 ^{o}c," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 292–301, Jan 2013.

[19] M. B. Kleiner, S. A. Kuhn, and W. Weber, "Thermal Conductivity of Thin Silicon Dioxide Films in Integrated Circuits," in *European Solid State Device Research Conference*, Sept 1995, pp. 473–476.

[20] P. R. Gray and D. J. Hamilton, "Analysis of Electrothermal Integrated Circuits," *IEEE Journal of Solid-State Circuits*, vol. 6, no. 1, pp. 8–14, Feb 1971.

[21] J. Angevare, L. Pedalà, U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "A 2800- μ m2 Thermal-Diffusivity Temperature Sensor With VCO-Based Readout in 160-nm CMOS," in *2015 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov 2015, pp. 1–4.

[22] K. Souri, Y. Chae, F. Thus, and K. Makinwa, "A 0.85V 600nW All-CMOS Temperature Sensor With an Inaccuracy of $\pm$0.4 ^{o}c (3σ) from -40 to 125 ^{o}c," in *International Solid-State Circuits Conference*, Feb 2014, pp. 222–223.

[23] G. Bosch, "A thermal oscillator using the thermo-electric (Seebeck) effect in silicon," *Solid-State Electronics*, vol. 15, no. 8, pp. 849–852, 1972.

[24] V. Szekely, "Thermal monitoring of microelectronic structures," *Microelectronics Journal*, vol. 25, no. 3, pp. 157–170, 1994.

[25] S. M. Kashmiri, S. Xia, and K. A. A. Makinwa, "A Temperature-to-Digital Converter Based on an Optimized Electrothermal Filter," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 2026–2035, July 2009.

[26] B. Vermeersch and G. De Mey, "A fixed-angle heat spreading model for dynamic thermal characterization of rear-cooled substrates," in *Twenty-Third Annual IEEE Semiconductor Thermal Measurement and Management Symposium*, March 2007, pp. 95–101.

[27] T. Veijola and M. Andersson, "Combined Electrical and Thermal Parameter Extraction for Transistor Model," *ECCTD*, pp. 754–759, September 1997.

[28] C. P. L. van Vroonhoven and K. A. A. Makinwa, "Linearization of a thermal-diffusivity-based temperature sensor," in *IEEE Sensors*, Oct 2009, pp. 1697–1700.

[29] AW Van Heerwaarden and PM Sarro, "Thermal monitoring of microelectronic structures," *Sensors and Actuators*, vol. 10, no. 3-4, pp. 321–346, Dec 1986.

[30] B. AW Van Heerwaarden, DC Van Duyn and PM Sarro, "Thermal monitoring of microelectronic structures," *Sensors and Actuators A: Physical*, vol. 22, no. 1-3, pp. 621–630, Jun 1990.

[31] S. M. Kashmiri, K. Souri, and K. A. A. Makinwa, "A Scaled Thermal-Diffusivity-Based 16 Mhz Frequency Reference in 0.16 μm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 7, pp. 1535–1545, July 2012.

[32] C. P. L. van Vroonhoven, D. d'Aquino, and K. A. A. Makinwa, "A thermal-diffusivity-based temperature sensor with an untrimmed inaccuracy of $\pm$0.2 ^{0}c (3s) from -55 ^{0}c to 125 ^{o}c," in *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, Feb 2010, pp. 314–315.

[33] T. Veijola, "Model for thermal spreading impedance in GaAs MESFETs," in *Proceedings of Third International Conference on Electronics, Circuits, and Systems*, vol. 2, Oct 1996, pp. 872–875 vol.2.

[34] T. I. T. R. for Semiconductors. ITRS 2007 report. [Online]. Available: http://www.itrs2.net/itrs-reports.html

[35] ——. ITRS 2011 report. [Online]. Available: http://www.itrs2.net/2011-itrs.html

[36] M. E. A. et. al., "Development of the self-aligned titanium silicide process for VLSI applications," in *IEEE Transactions on Electron Devices*, vol. 32, no. 2, Feb 1985, pp. 141–149.

[37] R. C. Jaeger, S. Hussain, J. C. Suhling, P. Gnanachchelvi, B. M. Wilamowski, and M. C. Hamilton, "Impact of mechanical stress on bipolar transistor current gain and Early voltage," in *IEEE SENSORS*, Nov 2013, pp. 1–4.

[38] X. Li, K. Maute, M. L. Dunn, and R. Yang, "Strain effects on the thermal conductivity of nanostructures," *Phys. Rev. B*, vol. 81, p. 245318, Jun 2010.

[39] M. A. Hopcroft, W. D. Nix, and T. W. Kenny, "What is the Young's Modulus of Silicon?" *Journal of Microelectromechanical Systems*, vol. 19, no. 2, pp. 229–238, April 2010.

[40] U. Sönmez, R. Quan, F. Sebastiano, and K. A. A. Makinwa, "A 0.008-mm2 area-optimized thermal-diffusivity-based temperature sensor in 160-nm CMOS for SoC thermal monitoring," in *European Solid State Circuits Conference*, Sept 2014, pp. 395–398.

3

Compact Phase Digitizers for Electro-Thermal Filters

This chapter gives an overview of readout architectures that digitize the temperature-dependent phase output of electro-thermal filters (ETFs). The area, accuracy, speed, resolution of these architectures will be considered. The chapter starts with the analysis of a completely analog gm-C based architecture followed by the analysis of a more digital VCO-based architecture. The circuit-level implementation of the two architectures is presented in chapters 4 and 5.

3.1. Introduction

Every physical sensor requires some kind of analog-to-digital converter to interface with the digital world. In the case of an electro-thermal filter (ETF), the desired conversion is from the phase (time) domain to the digital domain. When an ETF is to be used as a temperature sensor, an additional step is necessary to convert the digitized phase signal into temperature. As discussed in chapter 2, this phase-to-temperature property is related to the geometry and material properties of silicon and is very accurate once it has been characterized. The complete system, which will be referred to as a thermal diffusivity (TD) sensor, is shown in Fig. 3.1.

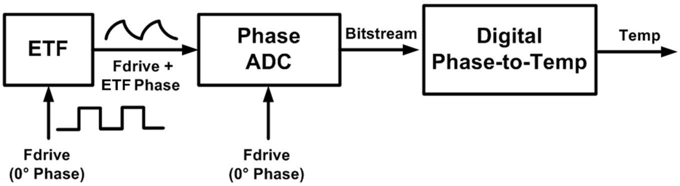


Figure 3.1: Block diagram of a TD sensor with an ETF, phase ADC and post-processing

In this chapter, we will focus on the system level implementation of phase ADCs for ETF readout. In particular, on the design of compact ADCs for thermal management applications. This chapter begins with a system level overview in section 3.2, which discusses the resolution and accuracy requirements of phase ADCs for ETFs.

Section 3.3 covers the fundamentals of phase detection and compares edge detection with coherent demodulation. Section 3.4 continues with the presentation and analysis of the Phase Domain ΣΔ Modulator (PDΣΔM), which is often used for ETF readout. The chapter then continues with sections 3.5 and 3.6, which discusses both Gm-C based and VCO-based PDΣΔM architectures. For improved performance, the former can be implemented as a two-step converter, while the latter can be implemented as a multi-bit converter. Drawbacks of the VCO-based architecture, such as time-domain quantization noise and counter wrap-around, are also discussed in section 3.6. The chapter concludes by summarizing the main characteristics of both Gm-C and VCO based PDΣΔMs. Their circuit-level implementations are presented in chapters 4 and 5, respectively.

3.2. System Overview

A practical phase ADC will have finite band-width and so will introduce phase error at certain frequencies. To analyze this error, we can split the ADC into a band-limited front-end and an infinitely fast phase detector as shown in Fig. 3.2. The front-end introduces an additional phase shift Φ_F, depending on its bandwidth BW. The spread in Φ_F will then contribute to the ADC's inaccuracy.

This spread will typically be a function of several circuit parameters (capacitance, transconductance, etc.) which will all spread independently. This makes Φ_F difficult to control, and hence the best strategy is to minimize it as much as possible. This leads us to our first requirement for good phase ADC accuracy: wide bandwidth.

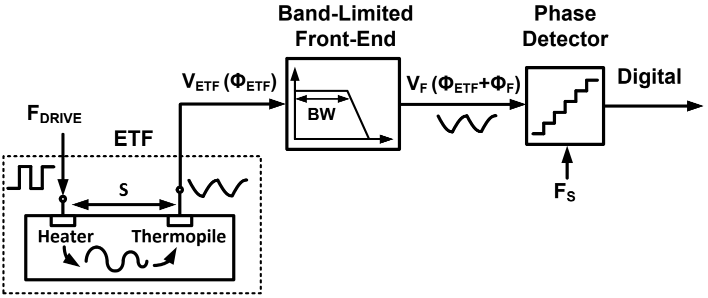


Figure 3.2: Block diagram of the TD sensor showing the phase ADC split into a band-limited front-end and phase detector

A phase ADC may also exhibit an offset Φ_{OS}, which can also degrade its accuracy. Again, this offset may spread due to process or temperature, and the best practice is to minimize it. The ADC's gain variations can also limit accuracy. However, this problem can be solved by the application of feedback in the phase domain. ADC nonlinearity is another source of error that can spread and introduce inaccuracy. However, it will be shown in section 3.6.4 that with proper design, the errors contributed by an even moderately non-linear ADC (~50 dB THD) will be negligible.

In a practical TD sensor, the dominant source of noise will be thermal noise, due to the mV-level output signals of the typical ETFs. The phase ADC's own noise should then be negligible compared to this. The low noise and wide-bandwidth requirements on the front-end require high power consumption, and thus the energy efficiency of the front-end itself also becomes critical. For these reasons, it is vital to choose a phase detection technique that gives the best signal-to-noise ratio for a given bandwidth or power. The following section discusses two common phase detection techniques and compares their energy efficiency.

3.3. Phase-to-Digital Conversion

Direct phase-to-digital conversion can be achieved in several ways, but in this section, only two methods will be discussed: edge detection and coherent demodulation. In edge detection, the timing information of a waveform is captured via threshold comparison or zero-crossing detection. In coherent demodulation, an input signal is multiplied by a sine- or square-wave signal at the same frequency to convert the phase information into a low-bandwidth, low-noise signal. Both methods are simple and require only a few circuit blocks to realize.

While edge detection seems to be a simpler solution, it is not suitable for the readout of noisy ETF signals. On the other hand, the accuracy of coherent demodulation depends on the stability of the ETF's output amplitude. However, this can vary greatly, e.g. due to variations in heater power and thermopile sensitivity. In the following sections, we will demonstrate how feedback can alleviate this problem. It will be shown that a ΣΔ architecture using coherent demodulation is an excellent way to achieve both accuracy and resolution.

3.3.1. Edge Detection

In edge detection, the phase of a periodic waveform can be extracted via the timing of its zero-crossing or mid-point voltage. This detection can be done by a comparator which compares the periodic signal to a threshold value. An example of edge detection is shown in Fig. 3.3, where V_S is the signal to be detected, and V_T is the threshold voltage. Timing instances t_0 and t_2 denote rising edges, while t_1 denotes a falling edge.

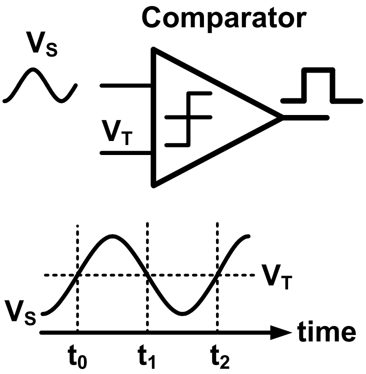


Figure 3.3: Timing diagram showing edge detection by a comparator

Intuitively, it can be seen that edge detection becomes more accurate if the timing instances t0, t1, and t2 are defined by sharper transitions. This means that a square-wave signal is preferred, which means a sharper derivative for V_S is desirable. To see how sharper rising/falling edges help, let's consider the case where a small error ΔV exists on V_S at the moments t_0, t_1 and t_2. Let's also define V_S as a sine-wave, as shown in Fig. 3.3:

$$V_S(t) = V_A sin(\omega t + \Phi) \tag{3.1}$$

Here, V_A is the amplitude, ω is 2π times the frequency, and Φ is the phase of the sine wave. The timing error on t0, t1 or t2 (Δt) can be defined as:

$$\Delta t = \Delta V \frac{dt}{dV_S} = \Delta V \frac{1}{\omega V_A cos(\omega t + \Phi)} \tag{3.2}$$

At the rising edge where the transition occurs, the cosine term assumes its maximum value. Therefore, we can take $cos(\omega t + \Phi)$ as 1, leading to:

$$\Delta t = \Delta V \frac{1}{\omega V_A} \tag{3.3}$$

As the amplitude of the signal increases, the timing error decreases as expected. Note that ΔV here is the instantaneous voltage error at the timing instances t_0, t_1 or t_2. This means that any noise in the signal will also be sampled by the comparator.

As discussed in Chapter 2, an ETF generates a small signal, typically a few mV peak-to-peak, along with a lot of thermal noise. As a result, a simple comparator may detect multiple crossings in a short period, severely impacting the inaccuracy and resolution of the system. To avoid this, a Schmidt trigger can be used. This replaces the single threshold V_T with two threshold values V_{T1} and V_{T2}, corresponding to rising and falling edges. The hysteresis provides reliability and noise immunity but may introduce spread in the exact values of V_{T1} and V_{T2}, which will increase inaccuracy. As a result, edge detection is not usually employed for ETF readout.

3.3.2. Coherent Demodulation

Coherent demodulation is a well-known technique for demodulating amplitude-modulated (AM) and phase-modulated (PM) signals. A coherent demodulator works by multiplying the input signal with a reference/carrier frequency. Fig. 3.4 shows the block diagram of a coherent demodulator.

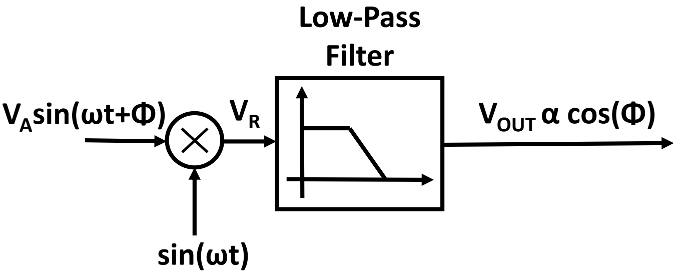


Figure 3.4: Block diagram of a coherent demodulator

The diagram shown in Fig. 3.4 is also known as a mixer or homodyne detector in an RF receiver chain. The input signal V_S is mixed with a sinusoidal reference V_D at the same frequency, and their product V_R is given as:

$$V_R = V_A sin(\omega t + \Phi) sin(\omega t) = 0.5 V_A cos(\Phi) + 0.5 V_A cos(2\omega t + \Phi) \tag{3.4}$$

When V_R passes through an ideal low-pass filter, the high-frequency cosine term disappears, leaving the DC term. V_D may also be a square wave signal with a fundamental harmonic amplitude of $\frac{4}{\pi}$, in which case V_R becomes:

$$V_R = \frac{2}{\pi} V_A cos(\Phi) + \frac{8}{3\pi} V_A cos(2\omega t + \Phi) + \frac{16}{15\pi} V_A cos(4\omega t + \Phi) + ... \tag{3.5}$$

The low-pass filter removes all the harmonics and leaves only the DC term. As long as the thermal noise superimposed on V_S has a larger bandwidth than F, the process of demodulation will not disturb its variance [3]. The noise bandwidth is defined by the low-pass filter, and thus the wideband noise component of V_S will be suppressed. One problem that remains is how to deal with the amplitude (V_A) term in equation 3.5. Since the amplitude of the input signal is often process dependent, we would like to get rid of this term.

3.4. Phase Domain ΣΔ Modulator

Coherent demodulation can be combined with feedback to cancel the amplitude component in equation 3.4. An example of this is shown in Fig. 3.5, with an ideal integrator and a voltage-controlled delay element (VCDL). The VCDL adds a phase Φ_D to the demodulation waveform, and the transfer function relating its input voltage to phase delay is defined as $K_{\Phi D}$.

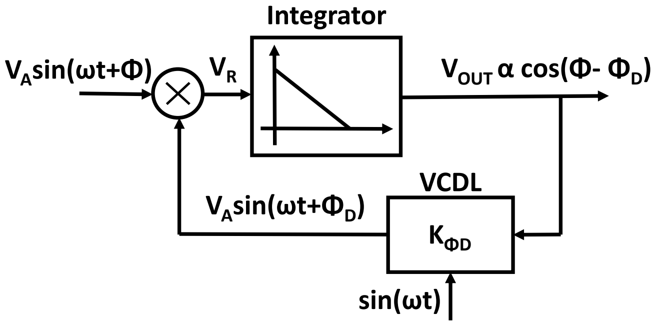

Integrator
V_Asin(ωt+Φ)
V_R
V_{OUT} α cos(Φ- Φ_D)
VCDL
V_Asin(ωt+Φ_D)
K_{ΦD}
sin(ωt)

Figure 3.5: Block diagram of a coherent demodulator with phase feedback

Since the integrator has infinite gain at DC, we get $V_R \approx 0$. This means that:

$$\Phi = \Phi_D - \pi/2 \tag{3.6}$$

And so, the output voltage is simply $(\Phi - \pi/2)/K_{\Phi D}$. We have removed the amplitude component at the output, but now a highly precise $K_{\Phi D}$ is required. This is a challenging problem for an analog delay element working with sine/cosine waves. However, it becomes much easier when the signal in the feedback path is a square wave. In the digital domain, we can then replace the VCDL with a Phase DAC, which delays a square wave in well-defined timing steps. These well-defined timing steps can be generated from a high-speed, accurate clock which is typically present in a CPU or SoC. In order to generate a direct digital output, an ADC can be included in the loop. The resulting system is shown in Fig. 3.6. The input is still a sine-wave, but the reference signal is a square wave with a phase of Φ_{DAC}. An M-bit quantizer is used to digitize the integrator output.

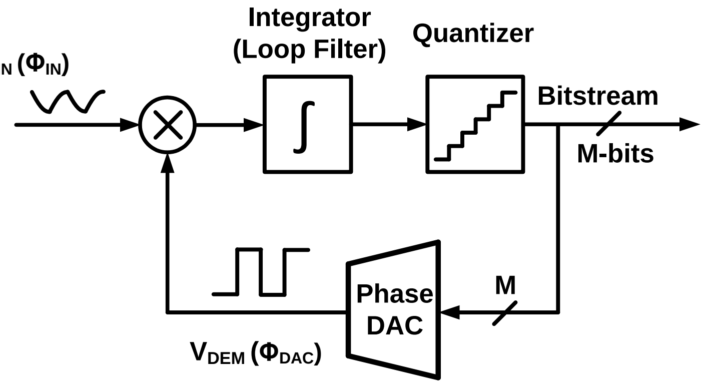


Figure 3.6: Block diagram of a simple phase-domainΣΔmodulator

The system in Fig. 3.6 is equivalent to a first-order ΣΔ modulator whose feedback is in the phase domain instead of the amplitude domain (voltage or current) [4]. It combines several architectural advantages:

- Good SNR due to the use of coherent demodulation (vs. edge detection)
- Good accuracy, since $K_{\Phi D}$ can be well defined in a digital DAC
- Direct digital output
- Simple architecture with easy-to-implement components (small area)

The main challenge with the architecture in Fig. 3.6 is the design of the input mixer, as its noise, timing accuracy, and signal isolation defines the performance of the system. Any crosstalk between its inputs will introduce a DC error and add inaccuracy. In the case of ETF readout, achieving this is particularly challenging since the ETF signal is very weak (mV-level), and the feedback signal is typically a rail-to-rail logic signal.

A practical way to ease this problem is to use an amplifier to boost the ETF signal. A good approach is to use a gm-stage (OTA) to convert the ETF voltage signal into a current, and use a simple chopper as the demodulator and a capacitor as the integrator. This gives rise to the Gm-C based phase domain ΣΔ modulator (PDΣΔM) described in the next section. The OTA's high DC gain and output impedance provide sufficient isolation between the feedback and ETF signals [4].

An important observation about the Gm-C based PDΣΔM is that it naturally derives from the simple, low-noise coherent demodulation technique with minimal additions. The addition of phase feedback is necessary to avoid amplitude dependence, while a quantizer is necessary to digitize the phase. A front-end is required to interface with the small ETF signal, and it also simplifies the integrator into a single capacitor. Thus, we have a simple architecture that is an excellent candidate for minimizing the readout area [4].

3.5. Gm-C Based PD ΣΔ Modulator

Fig. 3.7 shows the block diagram of a single-bit first-order Gm-C based PDΣΔM, along with the cross-section of an ETF. The ETF's voltage signal at frequency F_{DRIVE} is converted to current, and its phase shift is detected via the chopper driven by V_{DEM}. The resulting current is integrated over a capacitor and is fed into a latched comparator (sampled at F_S). The bit-stream output of the comparator switches F_{DEM} between two reference phases $\Phi 0$ and $\Phi 1$.

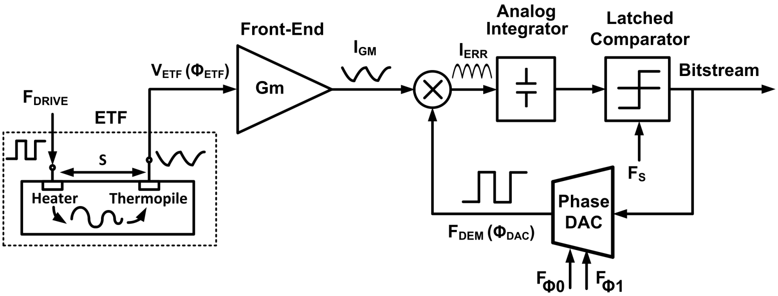


Figure 3.7: Block diagram of a single-bit, first order Gm-C based PDΣΔM

Since the average current over the integration capacitor is zero:

$$\mu cos(\Phi_{ETF} - \Phi_0) + (1-\mu) cos(\Phi_{ETF} - \Phi_1) = 0 \tag{3.7}$$

Then the average bit-stream value (μ) is related to the ETF and reference phases by [5]:

$$\mu = \frac{cos(\Phi_{ETF} - \Phi_1)}{cos(\Phi_{ETF} - \Phi_1) - cos(\Phi_{ETF} - \Phi_0)} \tag{3.8}$$

This is a non-linear function of Φ_{ETF}, but this so-called cosine non-linearity is quite systematic and so can be removed by post-processing [5]. However, since cos(x) is quite linear when x ~ 90 degrees, it becomes negligible for small phase DAC ranges ($\Phi_1 - \Phi_0$), e.g. for a 45 degree range the non-linearity is 10% while for a 5 degree range it is only 1.3%. To emphasize this, the phase references can be redefined with respect to a fixed offset of 90 degrees, or $\pi/2$ in radians. We define:

$$\Phi_0^{'} = \Phi_0 + \pi/2 \tag{3.9}$$

$$\Phi_1^{'} = \Phi_1 + \pi/2 \tag{3.10}$$

These definitions modify equation 3.8 into:

$$\mu = \frac{sin(\Phi_{ETF} - \Phi_1^{'})}{sin(\Phi_{ETF} - \Phi_1^{'}) - sin(\Phi_{ETF} - \Phi_0^{'})} \tag{3.11}$$

Here, we note that the function sin(x) approaches x when it is small. Therefore, when the differences between $\Phi_1^{'}$, $\Phi_0^{'}$ and Φ_{ETF} are small; equations 3.8 and 3.11 can be linearized. Thus, for small values of x we end up with:

$$\mu = \frac{\Phi_{ETF} - \Phi_1^{'}}{\Phi_0^{'} - \Phi_1^{'}} \tag{3.12}$$

Here, μ in equation 3.12 is a linear representation of Φ_{ETF} with respect to $\Phi_0^{'}$ and $\Phi_1^{'}$. It equals zero for Φ_{ETF}=$\Phi_0^{'}$, and one for Φ_{ETF}=$\Phi_1^{'}$. For other cases, the value of Φ_{ETF} can be calculated from μ with the knowledge of the references $\Phi_0^{'}$ and $\Phi_1^{'}$.

Chopper demodulation via V_{DEM} will introduce a strong tone at $2F_{DRIVE}$ (see eq. 3.5), which must be suppressed by the integrator to prevent possible intermodulation with the sampling frequency (F_S) or down-conversion of quantization noise into baseband. In practice, this means that a rather large integration capacitor must be used, which conflicts with the small-area target.

A more effective method is to use zero-crossing sampling [6] by ensuring that $F_S = F_{DRIVE}/K$, where K is an integer. This way, the system is synchronous with respect to both F_S and F_{DRIVE}. Hence, to the first-order, the error due to ripple between the sampling moments will not be visible at the output [6]. Figure 3.8 demonstrates the zero-crossing sampling technique with a ripple at F_{DRIVE} and sampling instances defined by $F_S = F_{DRIVE}/K$. In the figure, $K = 1$, and the DC voltage V_S is accurately sampled repeatedly, despite the large ripple. This way, we can tolerate a larger ripple and the integration capacitor size can be relaxed. A further constraint on F_S is that $F_S \leq F_{DRIVE}$, since the ETF's phase information is contained in one period of F_{DRIVE}. To achieve the lowest conversion time, we can select $F_S = F_{DRIVE}$ as in the figure.

For maximum accuracy, the modulator is usually operated as an incremental converter, in which the integrator is reset before each conversion [7]. A sinc filter (implemented by a simple counter) can be adopted for the decimation of the modulator's output [7].

3.5.1. Two-Step Conversion and Zoom ADC

The Gm-C based PDΣΔM has been successfully used in TD-based temperature sensors and frequency references [1][8]. Typically, a first-order incremental SDM is used to save area and guarantee accuracy. But this implies a long conversion time to achieve good resolution [8], since the quantization noise of the modulator can only be suppressed by increasing the length of the decimation filter, which implies slower conversion. Using a second order modulator can better suppress the quantization noise, but a second integrator is necessary to build it; which means

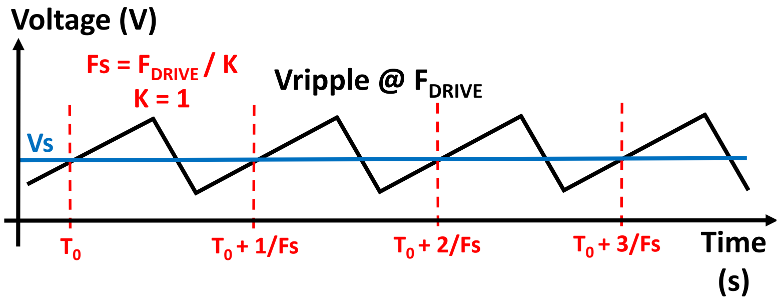


Figure 3.8: Demonstration of the zero-crossing sampling technique, with a ripple at Fdrive and synchronous sampling at a rate of Fs

an undesired increase in area. One solution to this problem is to use a two-step (coarse+fine) conversion, to realize a so-called zoom ADC [9].

This two-step conversion can be done by modifying the phase references $\Phi 0$ or $\Phi 1$ according to a coarse estimation of the input signal (ETF phase). The two phase references can be generated outside the TD sensors in a digital block (see section 4.3.4), just like the voltage references of an analog-to-digital converter. This is shown in Fig. 3.9, where some of the possible values of Φ_{DAC} are shown at the bottom. In this example, the phase DAC step (LSB) is 5.625° with a full span from 0 to 84.375°, and Φ_{ETF} is 60°.

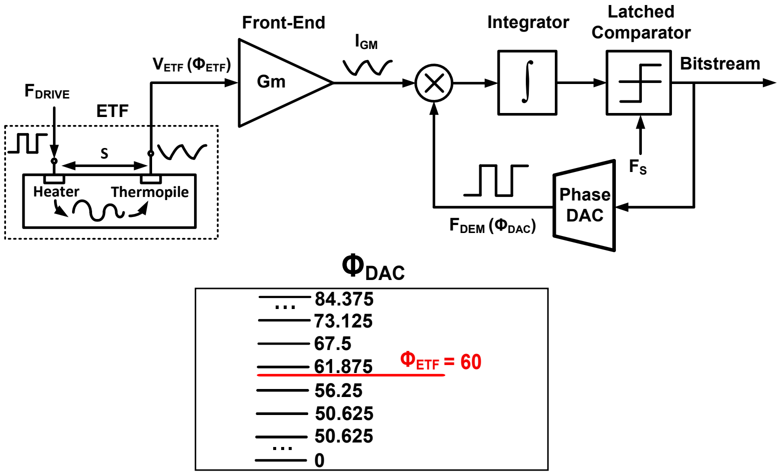


Figure 3.9: Block diagram of a Gm-C based PDΣΔM using two-step conversion. The phase DAC scale is shown at the bottom.

In the second phase, a longer fine conversion is made using the references $\Phi 0$ and $\Phi 1$. The reduced range reduces both quantization noise and integrator swing, and suppresses the systematic cosine non-linearity of the PDΣΔM. In contrast to residue-based two-step conversion techniques, no error from the coarse step is added to the fine conversion result, as long as the coarse estimation is accurate enough to enable the fine operation.

To guarantee this, references $\Phi 0$ and $\Phi 1$ are chosen to be $\pm$ 1 LSB away from the estimate instead of using the closest references ($\pm$ 0.5 LSB). This over-ranging guarantees that the fine conversion step will always straddle the input signal, as long as the coarse conversion error is less than half an LSB. As a disadvantage, due to over-ranging, the quantization noise of such a zoom ADC is double that of an ideal multi-bit PDΣΔM.

The coarse conversion has been previously done with a SAR [10] or ramp algorithm [9], but this requires extra logic and area. Alternatively, it can also be implemented as a short ΣΔ conversion. This mode of operation, first demonstrated in [11], is as follows: first, the extreme ranges of the ΣΔ references (0 and 84.375^oC in Fig. 3.9) are used during a short ΣΔ conversion. From this conversion result, the references closest to the input can be determined and then used during a longer, fine ΣΔ conversion.

In Chapter 4, the design of a prototype TD sensor using a Gm-C based PDΣΔM with two-step conversion is presented. It achieves an inaccuracy of 2.4 °C (untrimmed, 3σ) with a resolution of 0.2 °C (RMS) within 1 ms conversion time. The sensor area is only 8000 μm^2 in a mature 0.16 μm CMOS technology. As an important step, this sensor proves that compact and fast TD-based temperature sensors can be realized, albeit in older technology.

3.6. VCO Based PD ΣΔ Modulator

While the Gm-C based PDΣΔM is a good candidate architecture for compact readouts of ETFs, it does not inherently benefit from technology scaling in the same way as an ETF. In fact, both the gm-stage and the capacitors used in the Gm-C architecture are easier to implement in more mature technologies. The specific issues that limit the performance of Gm-C designs in advanced technologies are as follows:

1. The high DC gain required by the gm-stage can be hard to achieve because the intrinsic gain of transistors decreases in nm CMOS. Moreover, the supply voltage also decreases, creating headroom issues.

2. High-density capacitor area does not scale linearly with process. While metal or fringe capacitors scale (due to additional metal layers in the layer stack), scaled versions of MOS capacitors leak too much. Therefore, larger integration capacitors are required.

3. Similarly, the area of the gm-stage does not scale as well. Multi-stage or gain-boosted topologies are required to achieve sufficient DC gain, but this is at the expense of area.

For these reasons, the Gm-C architecture is not very suitable for use in advanced nm CMOS technologies. Figure 3.10 shows the Gm-C based PDΣΔM, highlighting the blocks that do not scale well. Note that all of these blocks are analog in nature.

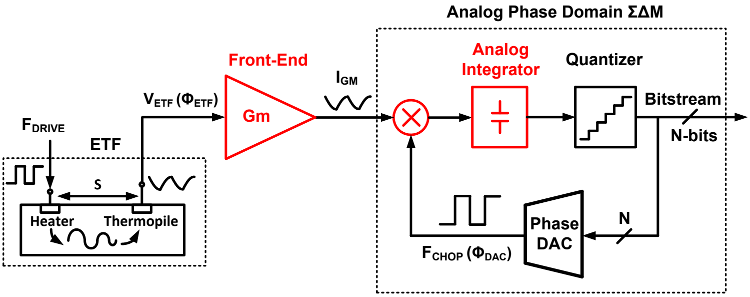


Figure 3.10: Gm-C based PDΣΔM with process scaling-averse blocks highlighted.

In order to design a more suitable architecture, let us look at the advantages of scaling down to an advanced process:

1. Faster operation of transistors due to higher f_T. For the same gm/Id ratio, transistor sizes and hence the capacitive load of each device can be reduced.

2. Smaller digital cells that consume much less power.

3. Potentially less parasitic capacitance in metal interconnects due to low-k dielectrics and thinner metal lines.

It seems clear that a more digital readout architecture would benefit greatly from technology scaling. However, implementing an accurate digital phase detector is quite challenging, especially when its area must be minimized.

One way to implement compact ADCs is by embedding a voltage-controlled oscillator (VCO) in a ΣΔ modulator [12][13]. In such architectures the VCO is used as an ideal integrator. This is because its output phase is an integral function of its control voltage [13]. Moreover, the phase output of the VCO can be easily digitized by using counters, and this leads to a highly-digital ADC architecture.

The simplest and smallest voltage or current controlled oscillator in CMOS is a chain of inverters, whose bias current is the control signal, as shown in Fig. 3.11. This is known as a ring oscillator. A cascade of N inverters (where N is odd and at least 3) will oscillate at a frequency F_{CCO} related to the input current I_{CCO}. F_{CCO} can then be digitally processed to obtain the original phase information. Since F_{CCO} is processed by digital cells, this architecture will take advantage of CMOS scaling.

In practice, the CCO cannot be directly connected to the ETF, so an intermediate gm-stage is necessary. The combined gm-CCO pair then behaves like a VCO.

In a VCO, the ETF's phase information will be frequency modulated (FM). To build a PDΣΔ loop, this information must be demodulated and integrated. Both these

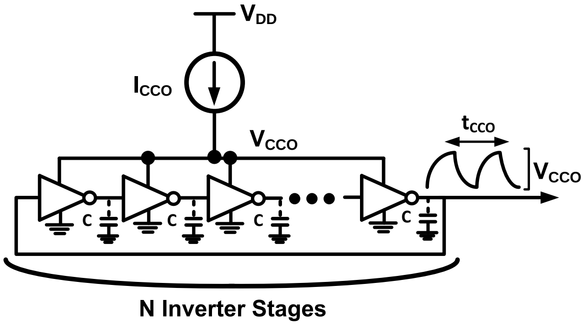


Figure 3.11: N-stage chain of inverters configured as a current controlled oscillator (CCO).

functions can be realized by an up/down counter [14]. The counter's up-down input ($CHOP$) facilitates chopper demodulation, i.e. multiplication by a square wave, since it determines whether the counter's state is either incremented or decremented. The value accumulated by the counter after one cycle of $CHOP$ will then be proportional to the integrated phase-shift between $CHOP$ and the VCO's output frequency, thus emulating the function of an integrator. As an example, suppose that the input to the VCO is a sine-wave with $I_{CCO} = Asin(\omega t)$, where A is the amplitude, and f is the frequency, and that the up/down signal is a square-wave that is shifted t_0 in time-domain with respect to I_{CCO}. After a single up count period of $0.5/f$, the number of counts C is given by:

$$C = \frac{K_{VCO}V_{IN}}{\pi F_{IN}} cos(\omega t_0) + \frac{f_{VCO}}{2f} \tag{3.13}$$

Here, ω is $2\pi f$, K_{VCO} is the VCO gain, and f_{vco} is the nominal VCO frequency. Here, the ideal demodulating square-wave ($CHOP$) is the up/down count signal that determines the phase shift t_0 but is otherwise locked to the frequency f. After the up and down periods are subtracted, the term $f_{vco}/2f$ is canceled, and only the cosine term that describes the phase shift between the envelope and up/down signals is left.

Coming back to the architecture, an M-bit register that samples the M MSB's of the counter's output is an efficient implementation of an M-bit quantizer. A multi-bit modulator can then be readily implemented since a very linear phase DAC can be realized with the help of an accurate reference clock [11][14].

With all of these modifications, the block diagram of the VCO-based modulator is as shown in Fig. 3.12. Compared to Fig. 3.10, the modulator uses more digital components, and is thus better suited for use in nm CMOS.

An implementation of a VCO-based PDΣΔM at block diagram level is shown in Fig. 3.13. An S-bit up/down counter performs both demodulation and integration, while an M-bit register acts as the quantizer. The up/down counter is a logic block that works on two asynchronous signals, and this means it can encounter metastability

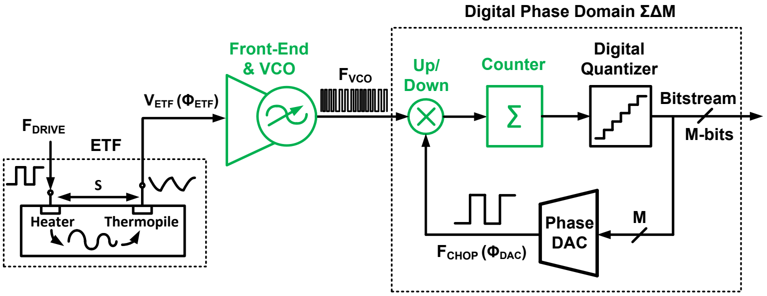


Figure 3.12: VCO based PDΣΔM with differences between Gm-C architecture highlighted.

problems when the UP/DOWN and VCO signals simultaneously trigger it. In order to prevent this, a flip-flop is used to synchronize the up/down signal with the next edge of the VCO clock. This is similar to the clock re-synchronization [15] required when two clock domains must cross each other.

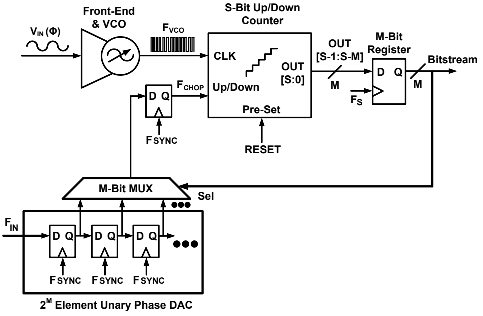


Figure 3.13: Block diagram of the implementation of a VCO-based PDΣΔM.

However, the VCO-based modulator also has some specific issues. First, it introduces additional quantization noise, since it only counts the edges of a frequency signal. The second issue is that the output of a counter will ”wrap around” rather than clip. These design problems will be tackled in sections 3.6.1 and 3.6.2.

In Chapter 5, a prototype TD sensor using 3-bit VCO based PDΣΔM in 40nm CMOS is presented. The architectural design method is described in this chapter, while the gm-stage and CCO design is presented in Chapter 5. Adopting a multi-bit VCO based PDΣΔM allowed the circuit to scale dramatically, down to an area of only 1650 μm^2. Thanks to improved lithography, the inaccuracy improves to 1.4 °C untrimmed (3σ), a nearly x2 improvement compared to [11]. It uses less power (2.5mW instead of 3.1mW), which results in a somewhat lower resolution of 0.36 °C (RMS) within 1 ms conversion time.

3.6.1. Time-Domain Quantization Noise

Unlike an analog integrator, an up/down counter can only count integer values and hence imposes rounding on its input. In the following analysis, a simple expression for the quantization noise due to the up/down counter will be derived.

For this analysis, we will model the counter as an ideal discrete-time integrator that introduces additional quantization error at its input at the end of every up/down cycle. The timing diagram in Fig. 3.14 shows how this simplification can be made. Here, we are also assuming that the input signal is a sine-wave with frequency F_{IN} and phase shift Φ_{IN} with respect to the reference square-wave up/down signal.

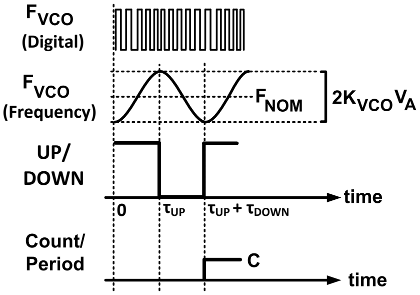


Figure 3.14: Timing diagram demonstrating how up/down counting can be modeled as a combination of chopping and discrete-time integration.

The frequency of the VCO (F_{VCO}) can be expressed as:

$$F_{VCO}(t) = K_{VCO}V_{IN}cos(2\pi F_{IN}t + \Phi_{IN}) + F_{NOM} \tag{3.14}$$

Here, K_{VCO} is the VCO gain, V_A is the amplitude of the input, and F_{NOM} is the nominal VCO output frequency. After integrating F_{VCO} for each full up period (τ_{UP}) and a full down period (τ_{DOWN}), an ideal counter, i.e. a counter without any quantization error, would compute the residual count C given by:

$$C = \int_0^{\tau_{UP}} F_{VCO}(t)dt - \int_{\tau_{UP}}^{\tau_{UP}+\tau_{DOWN}} F_{VCO}(t)dt \tag{3.15}$$

Every period, C is computed and then accumulated with the previous result. For an up/down signal with a duty cycle of 50% ($\tau_{UP} = \tau_{DOWN} = 0.5/F_{IN}$), C becomes:

$$C = \frac{-2K_{VCO}V_{IN}}{\pi F_{IN}} cos(\Phi_{IN}) \tag{3.16}$$

Shifting the phase of the up/down signal by Φ_{DAC} (due to the phase DAC action) is equivalent to shifting the input signal by - Φ_{DAC}; thus in a general form equation 3.16 becomes:

$$C = \frac{-2K_{VCO}V_{IN}}{\pi F_{IN}} cos(\Phi_{IN} - \Phi_{DAC}) \tag{3.17}$$

We define $\Phi'_{DAC} = \Phi_{DAC} + \pi/2$, which results in:

$$C = \frac{-2K_{VCO}V_{IN}}{\pi F_{IN}} sin(\Phi_{IN} - \Phi'_{DAC}) \tag{3.18}$$

For small values, $sin(\Phi_{IN} - \Phi'_{DAC}) = \Phi_{IN} - \Phi'_{DAC}$,and we can model the relationship between C and phase as a gain factor K (Fig. 3.15).

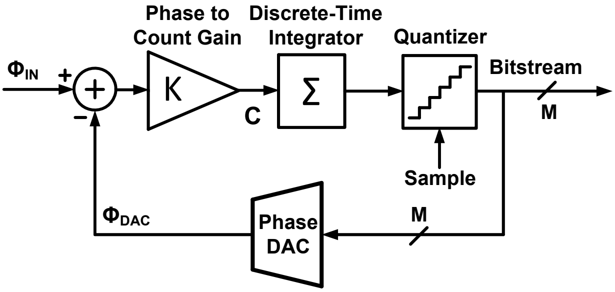


Figure 3.15: Block diagram of the ideal discrete-time PDΣΔM with a discrete-time integrator.

During regular ΣΔ operation, the feedback loop ensures that on average $sin(\Phi_{IN} - \Phi'_{DAC}) = 0$, thus validating our previous assumption. K or the phase-to-count gain (in degrees) can be readily defined from equation 3.18 as:

$$K = \frac{K_{VCO}V_{IN}}{90°F_{IN}} \tag{3.19}$$

However, a digital counter can only accumulate integer values because it only responds to the edges of F_{VCO}, which is equivalent to rounding C to an integer before the accumulation operation.

Fig. 3.16 demonstrates the timing diagram resulting from such synchronization. With this additional synchronization step, the quantization is in essence a truncation operation. The errors $\Delta Q_U(N)$ and $\Delta Q_D(N)$ denote the fractional count error at N^{th} up and down cycle, and as round-up errors they are bounded by [0 1] (Fig. 3.16).

$\Delta Q_U(N)$ and $\Delta Q_D(N)$ are deterministic for a given F_{VCO} and up/down signal. However, due to the significant dithering introduced by thermal noise, the quantization error can be assumed to be uniformly and randomly distributed on the [0 1] interval. This is analogous to approximating the quantization error introduced by the comparator of a ΣΔ modulator as white noise [17]. The variance of $\Delta Q_U(N)$ and $\Delta Q_D(N)$ is known to be [18]:

$$\sigma_Q^2 = \int_0^1 (q - 0.5)^2 dq = \frac{1}{12} \tag{3.20}$$

As can be seen on Fig.3.16, the total error for the Nth cycle ($\Delta Q_T(N)$) is given by:

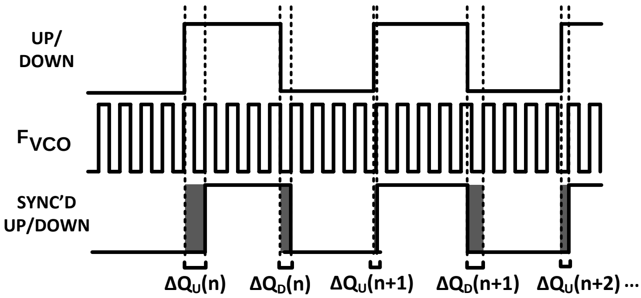


Figure 3.16: Timing diagram demonstrating the error introduced by metastability synchronization of up/down signal to F_{VCO}.

$$\Delta Q_T(N) = \Delta Q_U(N) - 2\Delta Q_D(N) + \Delta Q_U(N+1) \tag{3.21}$$

From equation 3.21, the variance of $\Delta Q_T(N)$ is $6\sigma_Q^2$. Note that the factor 6 arises from $\Delta Q_D(N)$ appearing in both up and down count operation. The total error after N up/down cycles can be written as the sum of the following series:

$$\sum_{k=1}^{N} \Delta Q_T(k) = \Delta Q_U(1) - 2\Delta Q_D(1) + 2\Delta Q_U(2) - 2\Delta Q_D(2)\ldots + \Delta Q_U(N) \tag{3.22}$$

Since each element in the series has a variance of σ_Q^2 and is uncorrelated from each other, the variance of the total error is equal to the sum of all component variances:

$$\sigma^2\left[\sum_{k=1}^{N} \Delta Q_T(k)\right] = (8N-2)\sigma_Q^2 \tag{3.23}$$

While the mean of the total error is zero. When $N \gg 2$, this error converges to $8N\sigma_Q^2$, and the expected error per integration cycle is $8\sigma_Q^2$. The bandwidth of this error is $F_{IN}/2$ since it manifests itself at the end of every complete up/down count period. Using equation 3.20, we get the total power of the error in fractional counts (σ_{TOTAL}^2) for a bandwidth F_{BW}:

$$\sigma_{TOTAL}^2 = \frac{2}{3} * \frac{F_{BW}}{F_{IN}/2} \tag{3.24}$$

If the sampling rate (F_S) of the PDΣΔM is chosen as F_{IN}, then the ratio $F_{IN}/2F_{BW}$ is equal to the oversampling ratio (OSR) of the ΣΔ modulator. Now, we can replace the discrete-time block in Fig. 3.15 with an additive white noise source (ΔQ_{ERR}) with a power of σ_{TOTAL}^2 to obtain Fig. 3.17.

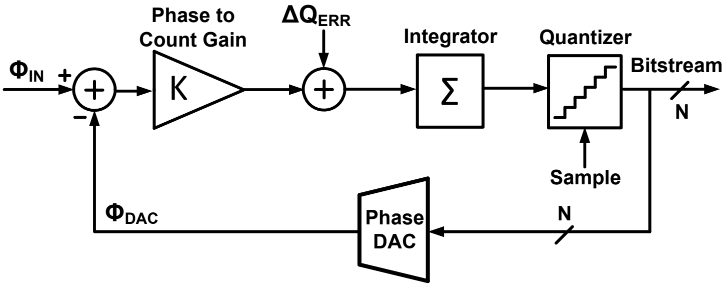


Figure 3.17: Block diagram of the PDΣΔM with a digital counter, with the quantizer replaced with a constant-power additive white noise source

The error in fractional counts can be directly converted into phase, which results in an input-referred phase error with an in-band power of σ_P^2, where:

$$\sigma_P^2 = \frac{2}{3OSR * K^2} \tag{3.25}$$

By using equation 3.19, the RMS in-band error in degrees ($\sigma_{P,^\circ}$) is:

$$\sigma_{P,^\circ} = \sqrt{\frac{2}{3OSR}}\frac{90^o F_{IN}}{K_{VCO}V_{IN}} \tag{3.26}$$

Some important conclusions can be derived from equation 3.26:

1. Time-domain quantization noise behaves like thermal noise since its power reduces linearly with oversampling ratio (OSR) and quadratically with V_{IN}.

2. Improving VCO gain (K_{VCO}) reduces quantization noise without any increase in SNR

3. Disregarding secondary effects, VCO nominal frequency (F_{NOM}) has no impact on quantization noise

4. For the same conversion time, increasing input/ETF drive frequency (F_{IN}) increases quantization noise; since noise power scales quadratically with F_{IN} and decreases only linearly with higher OSR.

5. For a given F_{IN}, OSR, and V_{IN}; K_{VCO} should be designed according to resolution and quantization noise specifications.

In order to suppress time-domain quantization noise, F_{IN} needs to be kept as low as possible while K_{VCO} should be increased. While it seems K_{VCO} can be increased arbitrarily, it can only grow as much as F_{NOM} since negative frequency values are not possible. However, this necessitates a faster and more power-hungry counter.

System-level simulations also confirm this analysis. Fig. 3.18 shows the power spectral density (PSD) of two models: an ideal ΣΔ modulator with additive noise as shown in Fig. 3.17 (simulated in Matlab), and a transient simulation in CppSim [19], where the whole circuit was operated in time-domain. The calculated quantization noise floor from 3.26 is shown as the dashed blue line.

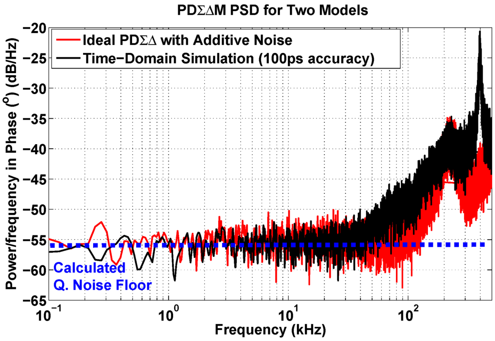


Figure 3.18: Power spectral density comparing the bitstream output of an ideal ΣΔ model with additive noise and transient simulations.

The block diagram of the CppSim model is shown in Fig. 3.19. A high-frequency clock (F_{SYNC}) is used to generate the 3-bit phase DAC values ranging from 11.25° to 90°. The up/down counter was compiled as a Verilog block and is hence ideal. Standard D flip-flop, VCO and multiplexer elements were used from CppSim's standard libraries.

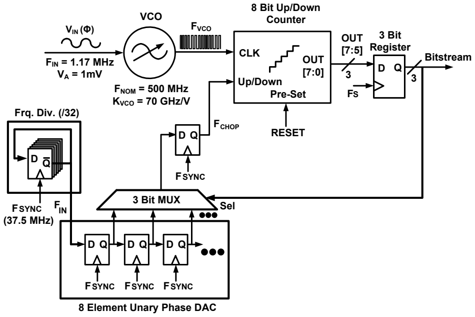


Figure 3.19: Block diagram of the implemented CppSim model.

In both models, $K_{VCO}V_{IN}$ = 70 MHz, F_{IN} = F_S = 1.17 MHz and F_{NOM} = 600 MHz. The phase DAC spans 78.75° with steps of 11.25°. The noise density of both models agree with each other at low frequencies, and also agree with eq. 3.26. The quantization noise is predicted to be 38 m° for OSR = 1024, which corresponds to ∼ 1ms conversion time. The difference between the models at high frequencies is thought to be due to the limited accuracy of the time-domain model and the related idle tone. The agreement of the two models with equation 3.26 means that long time-domain simulations are not required to determine the effect of time-domain quantization noise. Adding a simple noise source is sufficient to model its effects at low-frequencies.

3.6.2. Counter Size and Wrap Around

Due to practical limitations, the maximum counter output in a VCO-based PDΣΔM is limited, especially in compact readouts where the area of the counter must be minimized [14]. A possible issue is counter wrap-around, i.e. when the counter overflows. In order to establish the size of the counter, we will first investigate wrap-around. A straightforward solution would be to design the counter with overflow protection. Here, we will first observe what happens without any ΣΔ feedback, and both the input and the DAC phase are fixed. From equations 3.15 and 3.16, assuming equal up and down periods, we have the limitation on a non-wrapping counter size (C_{SIZE}) as:

$$C_{SIZE}(non-wrap) > \int_{0}^{\tau_{UP}} F_{VCO}(t)dt \tag{3.27}$$

Note that C_{SIZE} in this case must be at least larger than $F_{NOM}\tau_{UP}$, which is large (8 bits) for typical values (F_{NOM} > 500 MHz, τ_{UP} = 426 ns). A similar constraint also exists for the down counting phase.

If the counter is allowed to wrap (or overload) between up/down counts, this limitation is relaxed because only the remainder after up and down counting must be smaller than the counter size. This is expressed as:

$$C_{SIZE}(wrap) > C \tag{3.28}$$

where C is defined in 3.16, and does not ideally depend on F_{NOM}. Wrapping, in this case, means allowing the counter output to overflow, as shown in Fig. 3.20. Intuitively, eq. 3.28 means that the ΣΔ operates correctly if the counter wraps around during counting, as long as the output sampled by the quantizer is correct. This can be observed in Fig. 3.20, which shows how wrapping does not affect the latched counter result. Since a wrapping counter can be smaller and is easier to implement, we will not consider a counter with overflow protection.

The problem in a wrapping counter occurs when the counter value wraps around at or just before a sampling moment. This is illustrated in Fig. 3.21, which shows the sampled values of an 8-bit counter in a single-bit PDΣΔM. Due to ΣΔ modulation, the integrator output exercises a natural swing [14]. The counter wraps around when, for example, it is forced to go below a value of 0 due to this swing, instead of assuming a value close to the maximum. This corrupts both the current quantizer reading and the accumulated value in the counter.

For a single-bit PDΣΔM; wrap around can be avoided if the peak-to-peak swing of the latched counter value is less than half the counter length. In that case, the ΣΔ output bit-stream is the sampled counter MSB, and this length is 2^{S-1}, where S is the number of bits of the counter. Thus, we have:

$$C_{SIZE}(wrap) = 2^{S} > 2(C_{MAX} - C_{MIN}) = 2C_{PP} \tag{3.29}$$

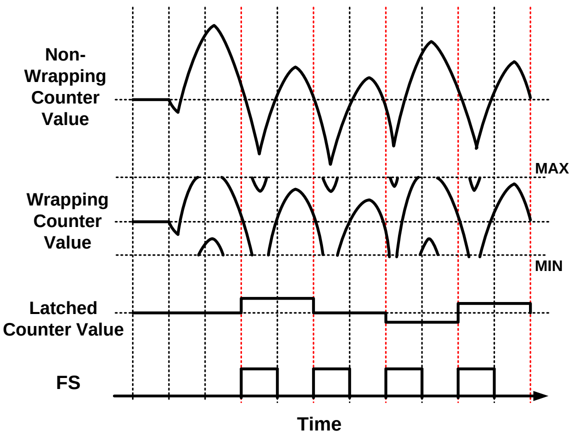


Figure 3.20: Timing diagram showing how a wrapping counter can tolerate a smaller swing and counter size.

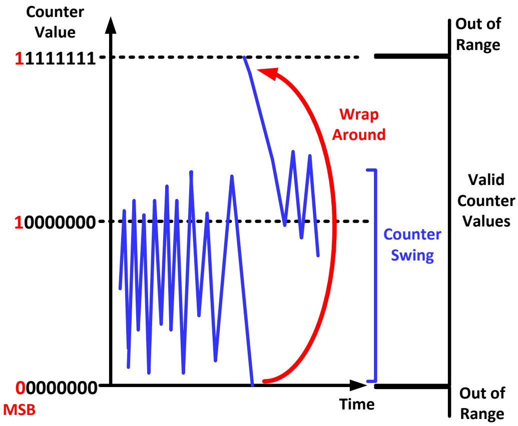


Figure 3.21: Latched counter values of a single-bit PDΣΔM, with an 8-bit counter, over time. If the counter wraps around, output is corrupted for several samples.

$$S > log_2(2C_{PP}) \tag{3.30}$$

Here, C_{PP} is the peak-to-peak swing of the counter. Since C_{PP} is dependent on the input signal amplitude (V_{IN}) and VCO gain (K_{VCO}), an interesting trade-off exists between counter size and quantization noise. For low quantization noise, V_{IN} and K_{VCO} need to be high (from eq. 3.26), which means a larger counter is necessary to avoid wrap-around. C_{PP} is bounded for a first-order modulator, and maximum swing is relatively constant over a bounded range of DC inputs [18]. Therefore, the behavior of counter swing can be simulated at a constant DC signal to assess C_{PP}. Then, counter size can be selected to avoid wrap-around. A similar analysis can be made for an M-bit PDΣΔM, as shown in Fig. 3.22. Counter wrap-around can be prevented if the peak-to-peak swing of the counter is guaranteed to be less than $2^{S\text{-}M}$. Therefore, for the general multi-bit case, we have:

$$C_{SIZE}(wrap) = 2^{S-M} > C_{PP}, forM > 1 \tag{3.31}$$

$$S > log_2(C_{PP}) + M \tag{3.32}$$

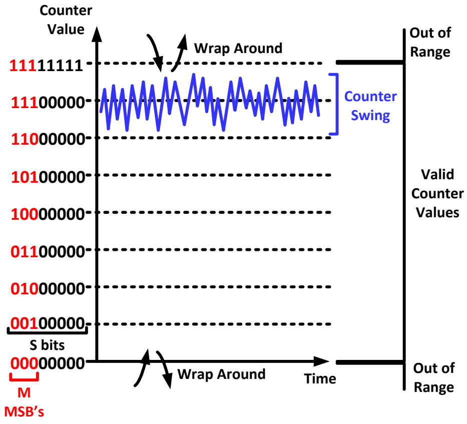


Figure 3.22: Counter values of a 3-bit PDΣΔM, with an 8-bit counter, over time.

As a convenient reliability measure, the input phase range can be restricted to be between the second highest and second lowest phase DAC steps, which relaxes the size of $2^{S\text{-}M}$ to only being larger than the peak-to-peak swing of the counter.

Another case where the counter can wrap around is at the start of the conversion after reset is released. This first count value can be too large and can cause a wrap around. We will assume that the input phase is bounded within [0 Δ], where Δ is the maximum value of the phase DAC (Φ_{DAC}), and the counter is reset to its median value ($2^{S\text{-}1}$). In this case, maximum value of $\Phi_{IN} - \Phi_{DAC} = \pm\Delta/2$, and from equation 3.18, we find the maximum count (C_{MAX}) to be:

$$C_{MAX} = \pm\frac{2K_{VCO}V_A}{\pi F_{IN}} sin(\Delta/2) \tag{3.33}$$

Intuitively, we can understand that C_{MAX} should not exceed half the counter length ($2^{S\text{-}1}$), or the counter will wrap around. This situation is similar to Fig. 3.22, except that the ΣΔ wraps around before the first sampling of the quantizer and shows potentially unstable behavior. Therefore, we have the restriction:

$$S > log_2\left(\frac{2K_{VCO}V_A}{\pi F_{IN}}sin(\Delta/2)\right) + 1 \tag{3.34}$$

Note that equations 3.33 and 3.34 hold true for a multi-bit PDΣΔM where the maximum initial error of $\Phi_{IN} - \Phi_{DAC} = \pm\Delta/2$. For a single-bit modulator, this error is ±Δ, and hence 3.33 and 3.34 must be modified by changing Δ/2 with Δ.

The constraint imposed by equation 3.34 is different from the one given by eq. 3.30 or 3.32, since it is not dependent on signal statistics. It is, instead, dependent on Δ, i.e. the span of the phase DAC. Resetting the counter to 2^{S-1} instead of to another arbitrary value also helps to minimize the counter size.

This theoretical analysis of wrap-around has also been verified with two simulation examples, for both single-bit and multi-bit cases. In both cases F_{IN} = 1.17 MHz as the input or ETF drive frequency.

The first example is a single-bit PDΣΔM with a Δ = 28.125°. V_{IN} is 1 mV and the typical K_{VCO} value is 110 MHz/mV; while the input referred noise density is 15 nV/√Hz. Fig. 3.23 shows the histogram of the simulated counter values for such a ΣΔ modulator with 8192 samples and a counter size of 8-bits. The peak-to-peak counter swing is 29 counts and from Eq. 3.29 S must be at least 6. Moreover, looking at equation 3.34, we get S > 5.8. In this case, both wrap-around conditions are averted if the counter is sized to be at least 6 bits.

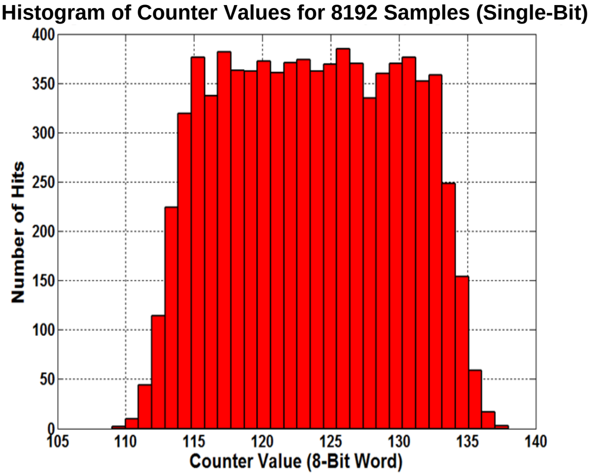


Figure 3.23: Histogram of counter values for 8192 samples for single-bit ΣΔ, 8-bit counter, phase range of 28.125°. Input is a 2 mVpp sine wave at 1 MHz with a phase shift of 58°.

The second example is the multi-bit (M = 3) PDΣΔM in [20], with a Δ = 78.75°. The histogram of the counter swing with an input amplitude of 0.65mV, K_{VCO} of 200 MHz/mV; and an input referred noise density of 15 nV/√Hz; is shown in Fig. 3.24. The peak-to-peak swing is 23 count values for 8192 samples. Thus, according to eq. 3.32, $S > 7$. Looking at equation 3.34, we get a more relaxed requirement of $S > 6.5$. The smallest counter size which avoids wrap-around for both conditions is 8-bits.

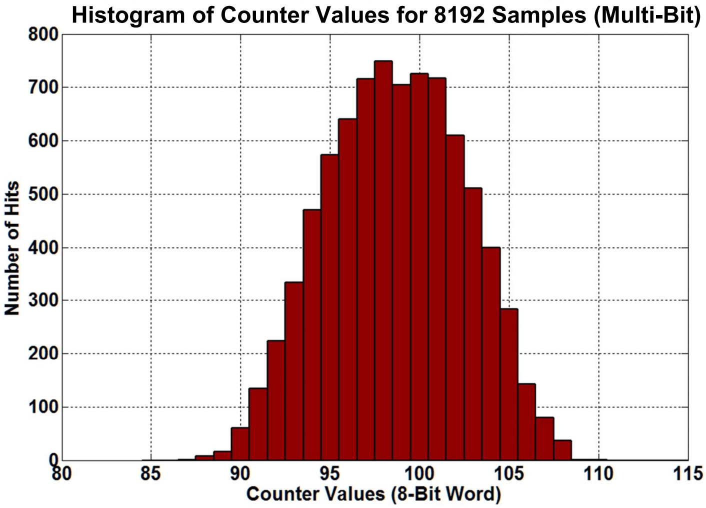


Figure 3.24: Histogram of counter values for 8192 samples for 3-bitΣΔ, 8-bit counter, phase range of 90°. Input is a 2mVpp sine wave at 1 MHz with a phase shift of 58.7°.

3.6.3. Multi-Bit Initial Reset and Settling

As also explained in the previous section, the counter in the multi-bit modulator must slowly settle at start-up, from the initial 2^{S-1} value to its nominal average value due to ΣΔ action. This is shown in Fig. 3.25, which shows the simulated settling of the model in Fig. 3.19 for a 70° input phase. Here, the counter was reset to zero instead of 2^{S-1} at the beginning of the conversion, to highlight the initial settling.

As can be seen in the figure, the first 100 samples of the bit-stream do not represent the correct phase value, and thus must be discarded. Thus, due to the settling time limitation, the conversion time slightly increases. The number of cycles that must be discarded (P) can be calculated by linearizing the behavior of the modulator. After this linearization, settling assumes the response as shown in Fig. 3.26. The settling error on the bit-stream at the n-th sampling cycle is denoted as $e(n)$ and is related to the previous state $e(n-1)$, and it can be shown:

$$e(n) = e(n-1)(1-r(n)) \tag{3.35}$$

Here, $r(n)$ can be understood as the normalized counter gain. It is defined as:

$$r(n) = \frac{C(n)}{e(n-1)} \tag{3.36}$$

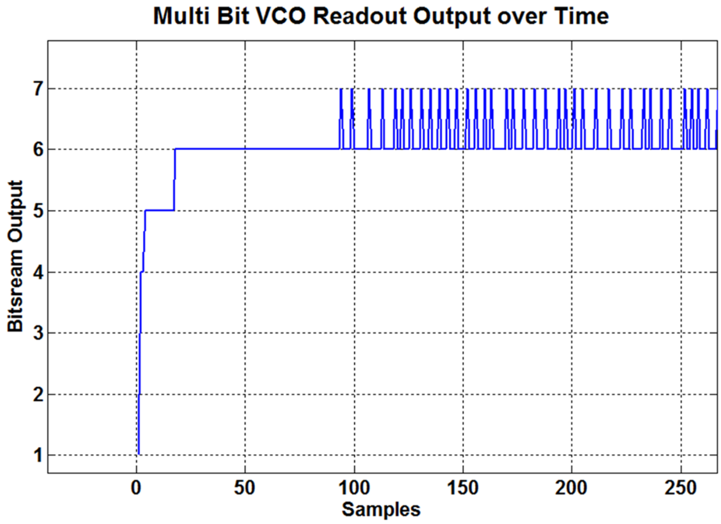


Figure 3.25: Bitstream output of a 3-bit PDΣΔM with 8-bit counter, right after resetting the counter to zero.

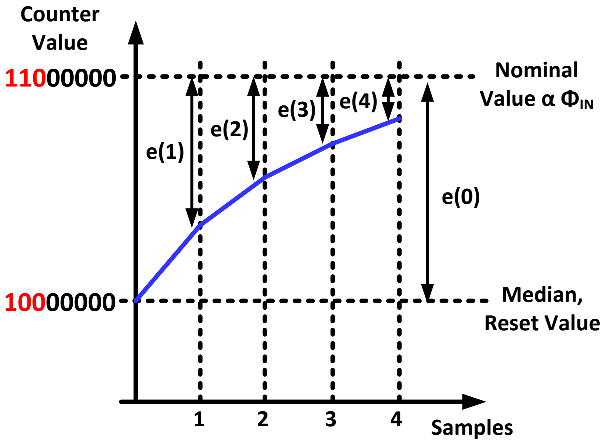


Figure 3.26: Timing diagram showing the linearized settling behavior in the multi-bit PDΣΔM.

where C, the number of counts, was defined before in equation 3.18. Simply, the gain term r expresses what percentage of the error is corrected at the next sample. Note the case when n=1 and $e(0)$ = 2^{S-1} :

$$r(1) = \frac{C_{MAX}}{2^{S-1}} \tag{3.37}$$

From equation 3.34, we know that r(1) < 1 to prevent wrap-around. Intuitively, this means that the modulator is allowed to operate only with an over-damped response. For simplicity, we will linearize $C(n)$, even though this assumption is not true for all input phases. If C is linearized around $(\Phi_{IN} - \Phi_{DAC})$, then using eq. 3.19, $r(n)$ becomes:

$$r(n) = \frac{K(\Phi_{IN} - \Phi_{DAC})}{e(n-1)} \tag{3.38}$$

The phase difference $(\Phi_{IN} - \Phi_{DAC})$ can be converted to a counter error since changes in counter MSBs linearly change the phase DAC value. For a phase DAC spanning Δ degrees, and an M-bit modulator, one LSB of the phase DAC corresponds to $\Delta/2^{M}$ degrees. The counter length that corresponds to this LSB is $2^{S\text{-}M}$. Thus, as expected, 2^{S} counter values correspond to Δ degrees. From this conversion, we get:

$$r(n) = \frac{K_{VCO} V_A \Delta (e_{IN} - e_{DAC})}{2^S F_{IN} 90° e(n-1)} \tag{3.39}$$

Where the difference $(e_{IN} - e_{DAC})$ is the error between the nominal counter value and the previous DAC value, or simply e(n-1). Thus, r simplifies to:

$$r(n) = \frac{K_{VCO} V_A \Delta}{2^S F_{IN} 90°} \tag{3.40}$$

As an example, we will calculate r for the parameters used in determining the quantization noise in Fig. 3.19. r = 0.23 for V_{IN} = 1mV, K_{VCO} = 70 MHz/mV, S = 8 and Δ = 90°; which obeys r < 1 condition to avoid wrap-around.

When r is constant, the solution to equation 3.35 is:

$$e(n) = e(0)(1-r)^n \tag{3.41}$$

The settling error $e(n)$ is a function of $e(0)$, or the initial difference between the input phase and the mean DAC value. Therefore, it is dependent on the input phase. In order to eliminate this input-dependent error, a certain number of cycles (P) must pass before conversion.

P can be calculated for a settling accuracy of A°, and for the worst case initial phase error of Δ/2:

$$P > log_{1-r}\left(\frac{A}{\Delta/2}\right) \tag{3.42}$$

P is a strong function of Δ and can impact accuracy. For r = 0.23, and Δ = 90°, 10 m° settling accuracy can be reached in under 32 samples; however, 68 samples must be skipped if the same level of accuracy is desired with half of the input signal amplitude, or half the K_{VCO}. It is also important to note that the linear analysis followed in this section is a rough estimation, and non-linear behavior of the modulator can significantly increase the practical limit imposed on P.

P, or the number of skipped bits, was then determined to guarantee an accurate result even if K_{VCO} or input signal amplitude decreases. In the intended application, the accuracy of the temperature sensor was paramount, so P was over-designed. Hence, although the phase range (Δ) was set to 78.75°, the initial counter value was set to 50° off from the input phase, higher than the Δ/2 = 39.375° worst case in reality. Input amplitude was also halved, to guarantee accuracy even if the signal is weaker than expected.

Then, r = 0.15 from equation 3.40 and for a settling target of A = 10m° , P > 52 from Eq. 3.42. This was validated by observing the bit-stream of the model in Fig. 3.19. Fig. 3.27 shows the settling of the modulator bit-stream during the first 128 samples. Even though it is hard to exactly determine the settling error for different values of P, it is clear that the result from Eq. 3.42 gives a rough but valuable estimation. P was chosen as 64 in the real implementation to allow the settling error to be much smaller than the resolution.

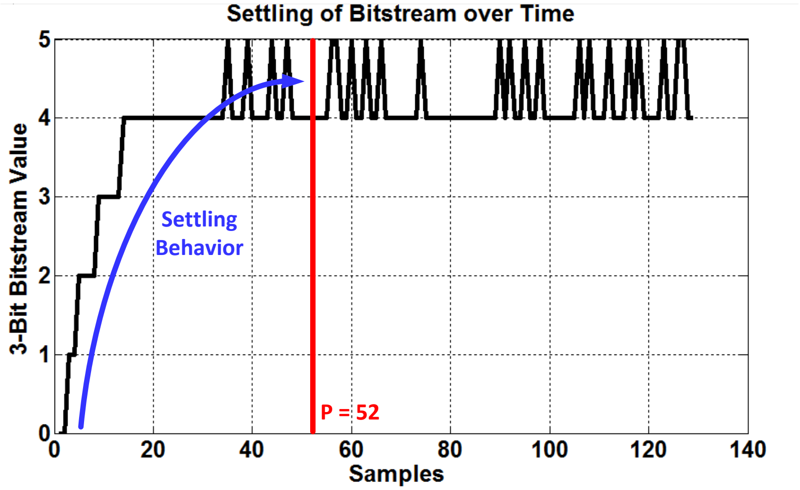


Figure 3.27: Settling of 50° initial error on a 3-bit PDΣΔM

3.6.4. Non-Linearity

Because of the cosine term in equation 3.18, the PDΣΔM exhibits systematic non-linearity. This non-linearity can either be corrected during digital post-processing [14], or by using small range(s) for Φ_{DAC} [5], which linearizes the sine term.

The following equation describes the non-linear relation between the average of output bit-stream (μ) and the input and DAC phase:

$$\mu = \frac{sin(\Phi_{IN} - \Phi'_{DAC,1})}{sin(\Phi_{IN} - \Phi'_{DAC,1}) - sin(\Phi_{IN} - \Phi'_{DAC,0})} \tag{3.43}$$

Here, $\Phi'_{(DAC,1)}$ and $\Phi'_{(DAC,0)}$ are the DAC phases for a feedback values of 0 and 1, with an additional $\pi/2$ phase shift as described in section 3.5. For multi-bit operation, $\Phi'_{(DAC,1)}$ and $\Phi'_{(DAC,0)}$ can be replaced with the exercised phase levels of the DAC. Note that μ itself is a non-linear function of Φ_{IN} but is not dependent on any circuit parameters, and its nonlinearity is thus systematic. However, the addition of another non-linear term can shift this systematic curve, and add inaccuracy to the design.

In the literature, the VCO is known to be a significant source of error if the signal is encoded in the voltage domain. For this reason, many techniques to compensate or cancel the VCO non-linearity are introduced in VCO-based ADCs [21][22][23]. However, since the information is encoded on the phase of the ETF signal, it will be shown that VCO non-linearity will have a smaller effect in a PDSDM. More importantly, while the following analysis is done for a VCO-based non-linearity source; it also applies to Gm non-linearity in a Gm-C based PDΣΔM.

For a sinusoidal input as in Eq. 3.14, the non-linearity of the VCO will produce tones at harmonics of F_{IN}. Considering only the second and third harmonic of $\omega_{IN} = 2\pi F_{IN}$, we get the non-linear frequency of the VCO as:

$$F_{VCO}(t) = A_1 cos(\omega_{IN} t + \Phi_{IN}) + A_2 cos(2\omega_{IN} t + 2\Phi_{IN} + A_3 cos(3\omega_{IN} t + 3\Phi_{IN}) + F_{NOM} \tag{3.44}$$

where A_N is the amplitude of the Nth harmonic component. Combining equations 3.18 and 3.45, we get the total count after up/down periods as:

$$C = \frac{-2}{\pi F_{IN}}[A_1 sin(\Phi_{IN} - \Phi'_{DAC}) + \frac{A_3}{3} sin(3\Phi_{IN} - 3\Phi'_{DAC}) \tag{3.45}$$

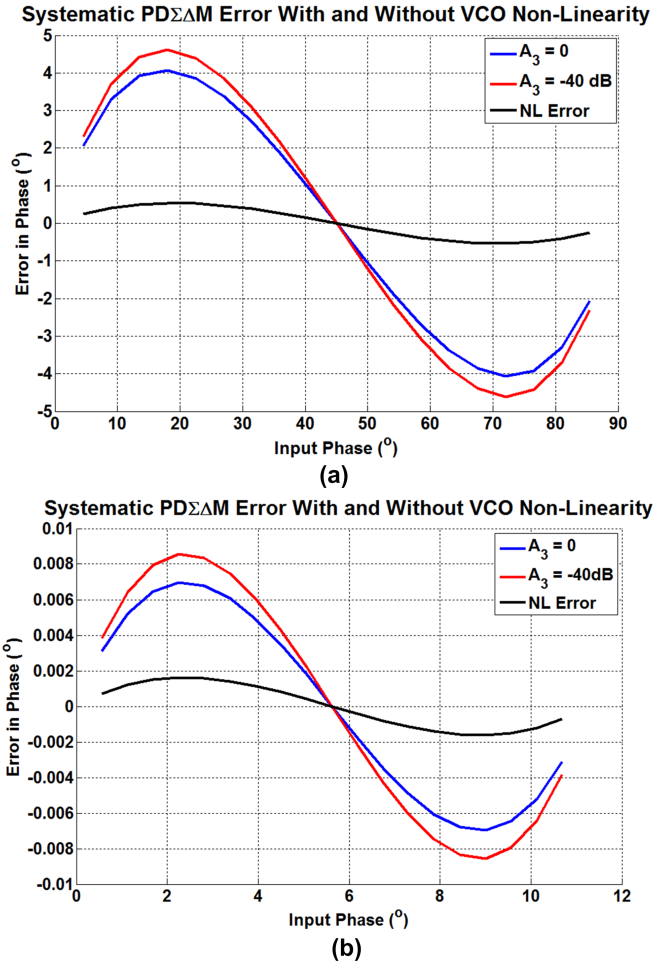
Due to the up/down operation, the second harmonic cancels out and third harmonic only adds a gain error as long as $3\Phi_{IN} - 3\Phi_{DAC}$ is small and $cos(3\Phi_{IN} - 3\Phi_{DAC})/3 \approx \Phi_{IN} - \Phi_{DAC}$. However, in the general case, equation 3.43 is modified to:

$$\mu = \frac{A_1 sin(\Delta\Phi_1) + \frac{A_3}{3} sin(3\Delta\Phi_1)}{A_1(sin(\Delta\Phi_1) - sin(\Delta\Phi_0)) + \frac{A_3}{3}(sin(3\Delta\Phi_1) - sin(3\Delta\Phi_0))} \tag{3.46}$$

where $\Delta\Phi_1$ is $\Phi_{IN} - \Phi'_{(DAC,1)}$ and $\Delta\Phi_0$ is $\Phi_{IN} - \Phi'_{(DAC,0)}$. In this case, the systematic nonlinearity is dependent on the gain terms A_1 and A_3, which depends on circuit parameters. If the ratio of A_3 to A_1 is fixed, the error can be eliminated by batch trimming. However, any spread on A_3 with respect to A_1 will add inaccuracy.

While this nonlinearity can appear to be a challenging problem to solve, the VCO is still linear enough in practical cases, and the additional error can be suppressed either by choosing a smaller phase range or multi-bit ΣΔ modulation.

As an example, we analyze the case where A_3/A_1 = - 40 dB and a single-bit modulator is spanning 90° range. 40 dB was chosen as a realistic number for the third-order nonlinearity of a typical VCO [12]. The systematic error of such a PDΣΔM over the full range is shown in Fig. 3.28 (a) as the red curve. The blue curve shows the case where $A_3 = 0$, and the black curve shows the difference between the two cases. The additional VCO nonlinearity causes a ±0.5° error; an error which is dependent on A_3, A_1 and input phase.



(a)

(b)

Figure 3.28: Non-linearity error of a PDΣΔM with and without the third order non-linearity introduced by the VCO with a phase DAC range of (a) 90° and (b) 11.25°

When the phase range is changed to 11.25°, as shown in Fig. 3.28 (b), the error then reduces to less than 2 m° or 180 ppm. This means that multi-bit or two-step PDΣΔMs [11] are relatively resistant to the nonlinearity of typical VCO designs, which exhibit third harmonic distortions of typically -40 to -60 dB. This is because both architectures use a finer phase DAC with smaller ranges.

3.7. Conclusions

In this chapter, several architectures for digitizing ETF signals have been described. The focus of the work in the chapter has been to simplify the readout architecture as much as possible, to save area in modern CMOS technologies. A secondary goal is to use digital rather than precision analog blocks that can scale well in such technologies. From an SNR point of view, it has been shown that coherent demodulation is better than edge detection for ETF readouts. Implementing feedback in a coherent demodulation scheme to improve accuracy naturally leads to the PDΣΔM architecture. Two different implementations of PDΣΔMs have also been described: Gm-C based (analog) PDΣΔM, and VCO-based (digital) PDΣΔM. The two-step conversion technique to further improve SNR, and save area in Gm-C based PDΣΔMs has been described.

The VCO-based PDΣΔM, which is a novel architecture for readout of ETFs, is described in the second part of the chapter. Using a VCO and a counter, it uses more digital components and thus allows the area of the readout to scale aggressively with the process node. Due to its compatibility with multi-bit feedback, it also makes two-step conversion redundant. However, it has several design challenges which must be solved, such as time-domain quantization noise, counter wrap-around, nonlinearity and settling time (for multi-bit modulators). In the end, the performance of VCO-based PDΣΔMs can be made good enough for thermal management applications by adopting the suggested design procedure.

References

[1] C. P. L. van Vroonhoven, D. D'Aquino, and K. A. A. Makinwa, "A thermal-diffusivity-based temperature sensor with an untrimmed inaccuracy of ±0.2°C (3σ) from -55 °c to 125 °C," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, Feb 2010, pp. 314–315.

[2] A. Demir, A. Mehrotra, and J. Roychowdhury, "Phase noise in oscillators: a unifying theory and numerical methods for characterization," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 5, pp. 655–674, May 2000.

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

[4] C. P. L. van Vroonhoven and K. A. A. Makinwa, "A CMOS Temperature-to-Digital Converter with an Inaccuracy of ±0.5°C (3 sigma)from -55 to 125°C," in *International Solid-State Circuits Conference*, Feb 2008, pp. 576–637.

[5] C. P. L. van Vroonhoven and K. A. A. Makinwa, "Linearization of a thermal-diffusivity-based temperature sensor," in *Sensors, 2009 IEEE*, Oct 2009, pp. 1697–1700.

[6] A. Bakker and J. H. Huijsing, "A cmos chopper opamp with integrated low-pass filter," in *Solid-State Circuits Conference, 1997. ESSCIRC '97. Proceedings of the 23rd European*, Sept 1997, pp. 200–203.

[7] J. Robert, G. C. Temes, V. Valencic, R. Dessoulavy, and P. Deval, "A 16-bit low-voltage CMOS A/D converter," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 2, pp. 157–163, Apr 1987.

[8] M. Kashmiri, M. Pertijs, and K. Makinwa, "A thermal-diffusivity-based frequency reference in standard CMOS with an absolute inaccuracy of ±0.1% from -55 °c to 125 °c," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, Feb 2010, pp. 74–75.

[9] C. van Vroonhoven, D. D'Aquino, and K. Makinwa, "A ±0.4 °C (3σ) -70 to 200 °C time-domain temperature sensor based on heat diffusion in Si and SiO2," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, Feb 2012, pp. 204–206.

[10] Y. Chae, K. Souri, and K. A. A. Makinwa, "A 6.3 μ W 20b incremental zoom-ADC with 6ppm INL and 1 μ V offset," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, Feb 2013, pp. 276–277.

[11] U. Sönmez, R. Quan, F. Sebastiano, and K. A. A. Makinwa, "A 0.008-mm2 area-optimized thermal-diffusivity-based temperature sensor in 160-nm CMOS for SoC thermal monitoring," in *European Solid State Circuits Conference (ESSCIRC)*, Sept 2014, pp. 395–398.

[12] G. Taylor and I. Galton, "A Mostly-Digital Variable-Rate Continuous-Time Delta-Sigma Modulator ADC," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 12, pp. 2634–2646, Dec 2010.

[13] M. Z. Straayer and M. H. Perrott, "A 12-Bit, 10-MHz Bandwidth, Continuous-time σ δ ADC With a 5-Bit, 950-MS/s VCO-Based Quantizer," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 805–814, April 2008.

[14] R. Quan, U. Sonmez, F. Sebastiano, and K. A. A. Makinwa, "A 4600 μ m2 1.5 °C (3 σ) 0.9kS/s thermal-diffusivity temperature sensor with VCO-based readout," in *Solid- State Circuits Conference - (ISSCC), 2015 IEEE International*, Feb 2015, pp. 1–3.

[15] R. B. Staszewski, C.-M. Hung, K. Maggio, J. Wallberg, D. Leipold, and P. T. Balsara, "All-digital phase-domain TX frequency synthesizer for bluetooth radios in 0.13 μ m CMOS," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, Feb 2004, pp. 272–527 Vol.1.

[16] C. C. Enz, E. A. Vittoz, and F. Krummenacher, "A CMOS chopper amplifier," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 3, pp. 335–342, Jun 1987.

[17] R. Gray, "Oversampled Sigma-Delta Modulation," *IEEE Transactions on Communications*, vol. 35, no. 5, pp. 481–489, May 1987.

[18] R. S. S.R. Norsworthy and G. C. Temes, *Delta-Sigma Data Converters*. Wiley-IEEE Press, 1997.

[19] M. H. Perrott. CppSim System Simulator Package. [Online]. Available: http://www.cppsim.com

[20] U. Sonmez, F. Sebastiano, and K. A. A. Makinwa, "1650 µm2 thermal-diffusivity sensors with inaccuracies down to ±0.75°C in 40nm CMOS," in *Solid- State Circuits Conference - (ISSCC), 2016 IEEE International*, Feb 2016, pp. 1–3.

[21] M. Park and M. Perrott, "A 0.13 μ m CMOS 78dB SNDR 87mW 20MHz BW CT Σ Δ ADC with VCO-based integrator and quantizer," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, Feb 2009, pp. 170–171,171a.

[22] K. Reddy, S. Rao, R. Inti, B. Young, A. Elshazly, M. Talegaonkar, and P. K. Hanumolu, "A 16mW 78dB-SNDR 10MHz-BW CT-$\Sigma\Delta$ ADC using residue-cancelling VCO-based quantizer," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, Feb 2012, pp. 152–154.

[23] A. Ghosh and S. Pamarti, "Linearization Through Dithering: A 50 MHz Bandwidth, 10-b ENOB, 8.2 mW VCO-Based ADC," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 9, pp. 2012–2024, Sept 2015.

4

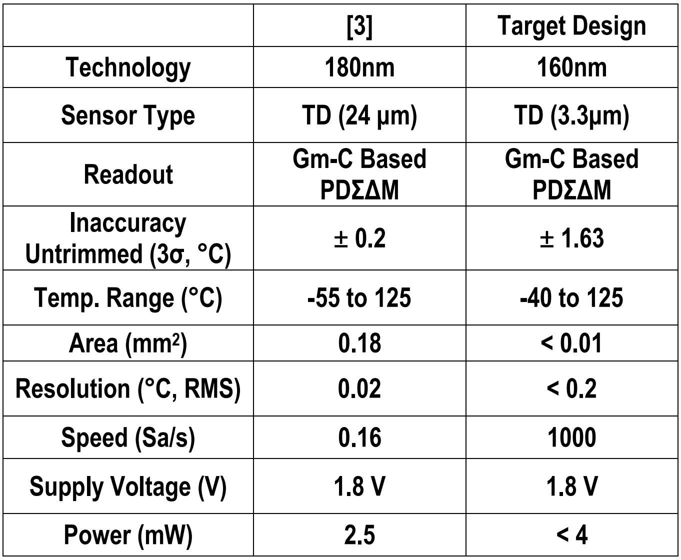
Area-Optimized Gm-C Based TD Sensors in 160nm CMOS

This chapter presents the design of an array of 8000 μm^2 1 kSa/s thermal-diffusivity (TD) temperature sensors in 160nm CMOS technology. They achieve an inaccuracy of ±2.4 °C (3σ) from -40 to 125 °C with no trimming, and ±0.65 °C (3σ) after a single-point temperature trim. They also achieve a resolution of 0.21 °C while dissipating 3.1 mW. This combination of accuracy, speed, and small size makes them well suited for thermal monitoring in microprocessors and other systems-on-chip. These results were achieved thanks to a simple but accurate front-end, scaled ETF design, and extensive use of digitally assisted analog design techniques.

4.1. Introduction

This chapter discusses the implementation of a scaled ETF and a compact PDΣΔM in a mature 0.16μm CMOS technology as a first step towards future implementations in deeper sub-μm technologies. The target specifications for the temperature sensor are shown in table 4.1, which compares them to the speed, power efficiency, area and expected inaccuracy of a previous TD sensor in a similar process node [3]. The aim is to reduce the area of the sensor, while dramatically increasing conversion speed and maintaining reasonable accuracy and resolution.

Table 4.1: Comparison table showing a previous, accurate state-of-the-art TD sensor and design targets in 0.16 μm CMOS



	[3]	Target Design
Technology	180nm	160nm
Sensor Type	TD (24 µm)	TD (3.3µm)
Readout	Gm-C Based PDΣΔM	Gm-C Based PDΣΔM
Inaccuracy Untrimmed (3σ, °C)	± 0.2	± 1.63
Temp. Range (°C)	-55 to 125	-40 to 125
Area (mm^2)	0.18	< 0.01
Resolution (°C, RMS)	0.02	< 0.2
Speed (Sa/s)	0.16	1000
Supply Voltage (V)	1.8 V	1.8 V
Power (mW)	2.5	< 4

To achieve this goal, the s3.3 ETF presented in Chapter 2 is combined with the compact PDΣΔM described in Chapter 3. Due to DRC rule restrictions, the s3.3 ETF is the smallest polygon ETF that can be implemented in the chosen 0.16μm technology.

The top-level design of the resulting TD sensor is described in section 4.2. This is followed in section 4.3 by a discussion of the circuit-level design of the PDΣΔM. Section 4.4 describes how multiple TD sensors are combined on the same die to realize a simple thermal management system. Section 4.5 reveals the sensor layout and top-level implementation. Further on, section 4.6 discusses the measurement results of the fabricated sensors and the chapter concludes with 4.7, which summarizes the performance.

4.2. System Level Design of a Single TD Sensor

As discussed in Chapter 3, the preferred phase digitizer for an ETF is an incremental PDΣΔM . Figure 4.1 shows the block diagram of a first-order incremental

PDΣΔM, as explained in section 3.5. Here, the Gm-stage and ideal multiplier functions are combined into a single Gm-demodulator; and a reset switch is added to facilitate incremental operation.

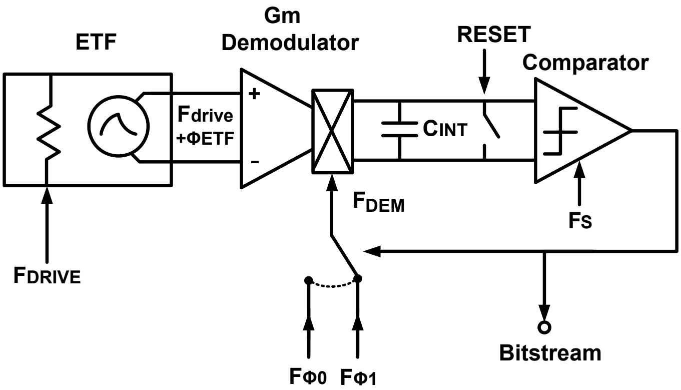


Figure 4.1: Block diagram of an Incremental PDΣΔM

In order to minimize area, a simple 1st-order modulator is preferred over higher-order modulators. However, the ratio of the optimum ETF drive frequency (around 1 MHz) to the target conversion rate of 1 kSa/s results in an oversampling ratio (OSR) of only 1000. With a target temperature range of -40 to 125 °C, the resulting quantization error is about 0.2 °C, which is of the same order as the ETF's thermal noise floor. In order to reduce the quantization error, the two-step conversion technique, as discussed in section 3.5.1 can be used.

Two-step conversion reduces both the systematic cosine non-linearity of the PDΣΔM as well as other circuit-related non-linearities (see sections 3.5.1 and 3.6.4). It also reduces the integrator output swing for a given integration capacitor, thus allowing a smaller capacitor to be used and improving the sensor's area efficiency. As a result, the Gm-stage can be implemented with a smaller and more energy-efficient telescopic amplifier, rather than the folded cascode used in [2].

4.3. Circuit Design

Since the ETF signal is very small (typically a few mV peak-to-peak), cross-talk from the large square-wave F_{DRIVE} applied to the heater or the input offset at the Gm-stage can corrupt the sensor's accuracy. The offset of the gm-stage is particularly problematic, since it is typically larger than the ETF signal and results in a large ripple at the demodulator output. Two techniques are implemented in the readout to suppress offset: system-level chopping and auto-zeroing, which is explained further on in section 4.3.1.

The problem of electrical cross-talk in an ETF is demonstrated in Fig. 4.2, which shows the top level layout and equivalent circuit schematic of a polygon ETF along with its heater-to-thermopile parasitics. The parasitics C_{p+} and C_{p-} create a high-pass path linking the signal on the heaters, typically a large square-wave, to the

ETF output. As a result, large spikes occur in output (V_+-V_-), thus adding phase errors.

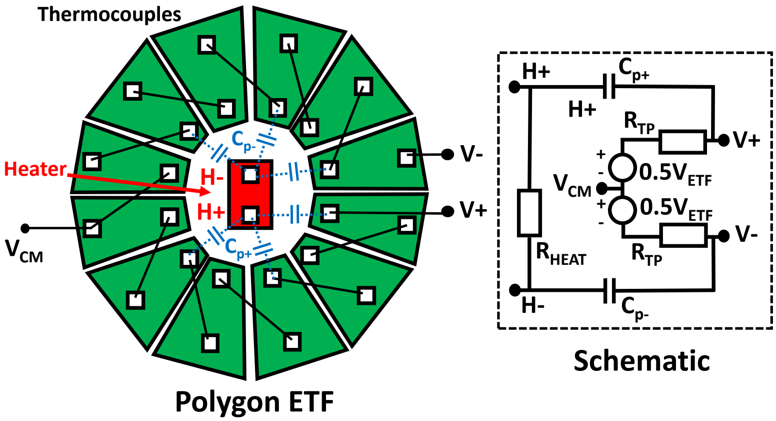


Figure 4.2: Top level layout of a polygon ETF, and its schematic equivalent circuit with parasitics between heater and thermopiles highlighted

One possible solution, discussed in [4], is to periodically invert the polarity of the heater's drive voltage to average the electrical crosstalk while preserving the desired heat signal. This so called heater-drive inversion (HDI) technique is shown in the timing diagram of Fig. 4.3, where the signals H_+ and H_- are shown along with the cross talk on nodes V_+ and V_-, as well as the power dissipated on the heater (P_{heat}). For the first two drive periods, the net electrical cross-talk (V_{xtalk}) has a positive phase error with respect to P_{heat}, while its polarity is inverted during the next two periods, resulting in a negative phase error. This technique ensures that the average phase error approaches zero.

System-level chopping is applied to the entire modulator by toggling F_{LOW} and digitally inverting F_{DRIVE} and BS. The final result is then the average of the two conversions. To facilitate incremental conversion, C_{INT} is reset between these two conversions to reset the memory of the modulator. Similarly, HDI is applied via F_{HDI}, which toggles the polarity of the electric pulses on the heater while maintaining the polarity of the pulses in the thermal domain [4]. HDI can be efficiently split up into two conversions, which results in four conversions when combined with system-level chopping. Adopting both low frequency chopping and HDI, we arrive at the system-level block diagram of a single TD sensor, along with its timing diagram, in Fig. 4.4.

With the two-step conversion approach, each conversion consists of a coarse and a fine step. These conversions are further split into four states, which are associated with system-level chopping and HDI. In the end, the results of each sub-conversion are averaged by in the modulator's decimation filter to arrive at the final result.

Each conversion begins by auto-zeroing the gm-stage and resetting C_{INT}. Dur-

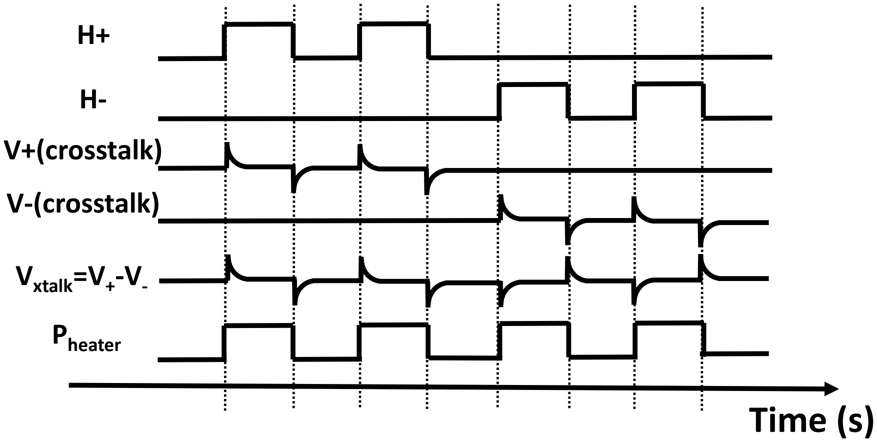


Figure 4.3: Timing diagram of heater nodes and electrical cross-talk visible at the ETF output, with heater drive inversion

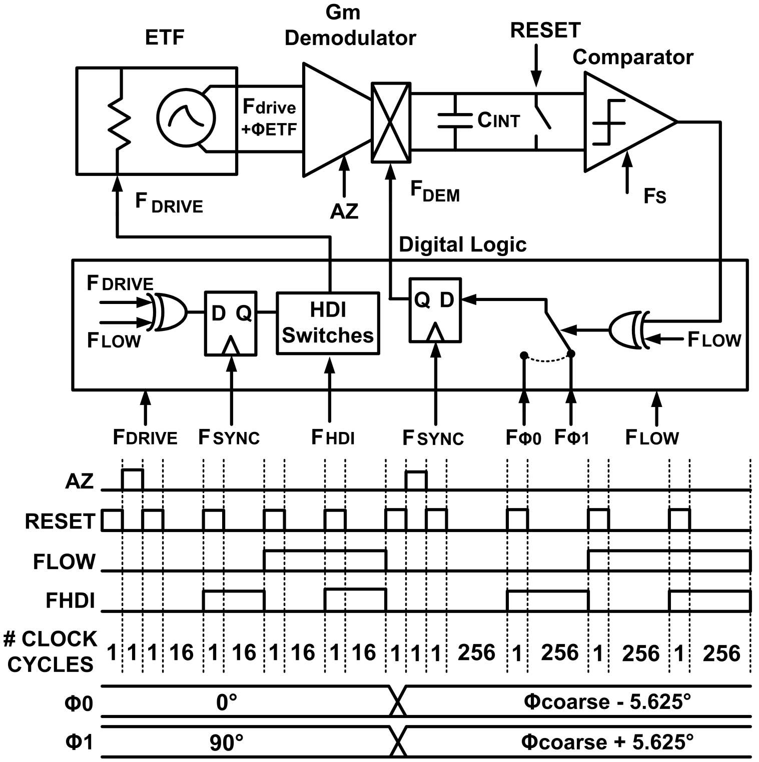


Figure 4.4: Expanded block diagram of an individual temperature sensor, along with the timing diagram

ing the coarse conversion, Φ_0 and Φ_1 are set to 0° and 90°, respectively, to cover the full temperature range, and a 64-step conversion is used to make a 4-bit estimate of the ETF's phase (Φ_{coarse}). In contrast to [1], in which the coarse conversion was based on a single-slope ADC, here a coarse ΣΔM was used because of its compatibility with chopping and HDI. During the fine conversion, Φ_0 and Φ_1 are set to straddle Φ_{coarse}, such that $\Phi_1 - \Phi_0$ = 11.25°. A 1024-step conversion is then used to accurately determine the ETF's phase shift (Φ_{fine}).

To minimize circuit delay errors, $F_{\Phi 0}$, $F_{\Phi 1}$ and F_{DRIVE} are synchronized to a reference clock F_{SYNC}, which is also used to generate the reference delays Φ_0 and Φ_1 in section 4.3.3. A full conversion then requires 1130 clock samples, resulting in a conversion rate of 1.04 kSa/s for F_{SYNC} = 75 MHz and F_{DRIVE} = 1.17 MHz.

The fine conversion range was limited to 11.25° in order to limit F_{SYNC}, since F_{SYNC} must increase if we want to reduce fine conversion range. The choice of 75MHz was motivated by its ready availability as a common XTAL/MEMS clock reference frequency. Moreover, it was shown in section 3.6.4 that a range of 11.25° results in a negligible cosine non-linearity.

4.3.1. Gm Demodulator

The gm stage, shown in Fig. 4.5, employs a compact telescopic topology that drives an integration capacitor C_{INT}. Due to the robustness of PDΣΔM to non-linearity, as described in section 3.6.4, and the reduced integrator reduced swing due to two-step conversion; C_{INT} (5.5 pF) can be realized as an area-efficient MOS capacitor. Because of area constraints, the area of the transistors in the gm-stage was minimized, which significantly exacerbates its offset. After chopping, this large offset (roughly 10 mV, 3σ) would create a large ripple voltage on C_{INT}. This increases the output voltage swing and might cause clipping. As shown in Fig. 4.5, auto-zeroing (AZ) was used to mitigate this ripple, by using an auxiliary gm-stage and capacitors to store the OTA's input-referred offset when the AZ signal is high. The AZ capacitors C_{AZ} (1 pF) were implemented as MOS capacitors to save area. The simulated residual offset after AZ is less than 10 µV, for the chosen 1 pF value for AZ capacitors. This is small enough that the resulting ripple is negligible compared to the modulator's quantization error.

The gm-stage employs gain boosting to improve its DC gain and mitigate the offset associated with its chopper demodulator [2]. Voltage headroom constraints force the input pair M7-8 to work in weak inversion and so, to minimize area, near-minimum length transistors were used. The resulting low output impedance leads to output offset current due to the offset of the cascode transistors (M5-6) [2]:

$$I_{OUT} = 4F_{DRIVE}C_{PAR}V_{OS5-6} + \frac{2V_{OS5-6}}{R_{O7-8}}$$

Here, I_{OUT} is the OTA output offset current, F_{DRIVE} is the chopper frequency, C_{PAR} is the parasitic capacitance seen at the drain of M7-8, V_{OS5-6} is the threshold mismatch of devices M5-6, and R_{O7-8} is the output impedance of M7-8. Simple pseudo-differential gain boosters, as shown in the figure, are sufficient to suppress this offset current without significantly affecting the area. Similarly, the PMOS cas-

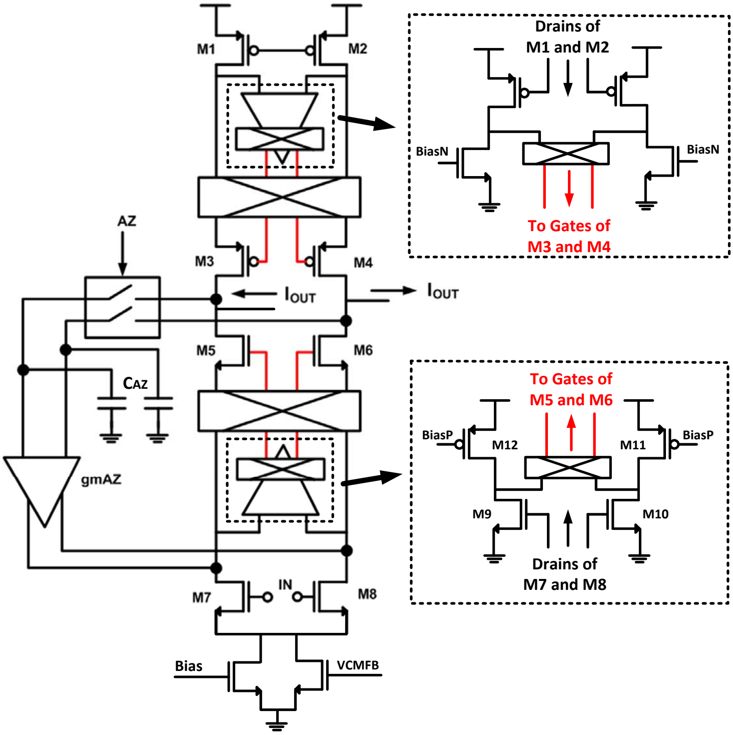


Figure 4.5: Circuit level implementation of the gain-boosted telescopic cascode gm-stage with auto-zeroing

codes M3-4 were gain boosted to limit the output offset current induced by their mismatch. Such gain boosters are commonly used in high-gain amplifier or current mirror circuits. The reduction in headroom at the output associated with the use of the common-source gain boosters is not a significant concern, as the output swing during fine conversion is low due to the adoption of the two-step ADC architecture. Any residual offset after gain boosting is canceled by system-level chopping. Simulations show that after AZ and system-level chopping, the error due to residual offset is less than 0.05 °C.

The Gm-stage is biased to operate with a DC current of 100 μA, with a transconductance (gm) of 1 mA/V. The input devices M7-8 are designed to work in weak inversion to achieve a gm/Id ratio of $\sim$ 20. Biasing of M7-8 is made PTAT to keep gm relatively constant over temperature. A common-mode feedback sense element in the comparator pre-amplifier detects the common mode output of the Gm-stage and corrects the biasing current to set 0.9V output common mode. With the s3.3 ETF connected, the noise floor of the gm-stage is 14 nV/$\sqrt{Hz}$, which results in a phase noise-density of 1.3 m°/$\sqrt{Hz}$ for a 2 mVpp ETF signal.

The operation of the gm-stage was simulated over proces corners and temperature. The worst-case DC gain over process corners and temperature is 88.5 dB as shown in Fig. 4.6. Given gm ∼ 1mA/V, this corresponds to 26 MΩ output impedance, which is high enough to suppress DC leakage current over C_{INT} due to ΣΔ swing.

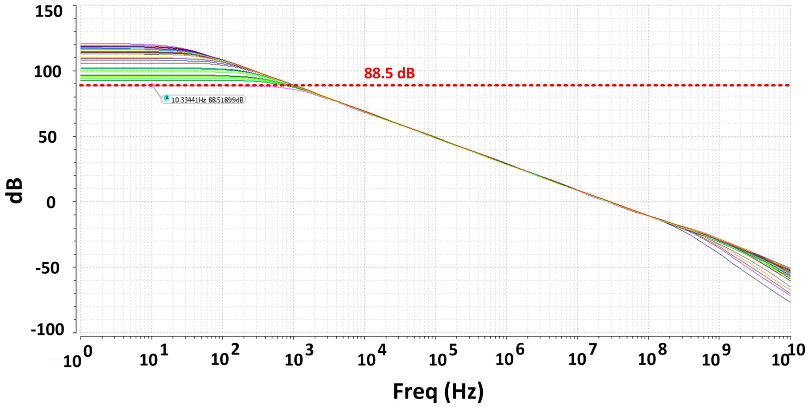


Figure 4.6: DC gain of Gm-stage over temperature and corners

The typical swing over C_{INT} is only ∼ 50 mV during fine conversion. This was inferred via transient simulation, which reveals the swing that can be tolerated before the integrator approaches saturation. Fig. 4.7 shows the integrated output of the OTA for a reference 11.25^o phase shift and C_{INT} of 5.5 pF. As the OTA integrates the phase error, the output swing increases over time due to integration. At some swing level, the OTA cannot accommodate more voltage swing, and integration stops. Over corners and temperatures, the worst case tolerable swing over C_{INT} is ∼ 150-250 mV; which is much higher than ∼ 50 mV swing due to fine conversion.

The limited swing at the Gm-stage output also causes non-linearity before approaching saturation. As was shown in section 3.6.4, this non-linearity is not a significant problem for a PDΣΔM. It was shown that a third-order non-linearity in the order of -40dB or 1% results in only $2m^o$ error. Thus we assume that the modulator behaves normally until the Gm-stage approaches saturation as discussed above.

4.3.2. Comparator

Figure 4.8 shows the schematic of the dynamic comparator with pre-amplification. A pre-amplifier (M1-4) was used to minimize comparator kickback and offset (M7-11). Comparator offset can constitute a significant error during the short coarse ΣΔ conversion since the noise transfer function (NTF) of a 1st-order incremental ΣΔ modulator is limited by the number of conversion cycles (or OSR). For a 1st-order PDΣΔM conversion consisting of N periods, the integrated voltage at the comparator input can be approximated as:

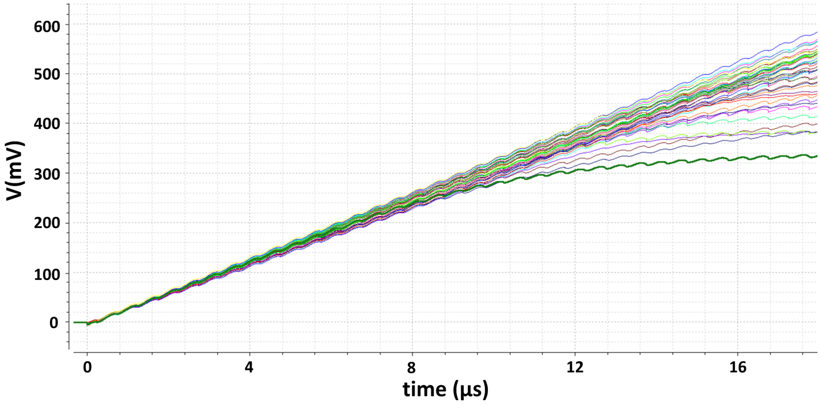


Figure 4.7: Integrator swing over C_{INT} over corners and temperature

$$V_{INT} = \frac{A_{ETF} g_M N sin(\Phi_{ETF} - \Phi_{0,1})}{2 C_{INT} F_S} \tag{4.1}$$

Here, A_{ETF} is the ETF signal amplitude, $\Phi_{0,1}$ is the mean phase applied via the PDΣΔM and the phase $\Phi_{ETF} - \Phi_{0,1}$ is the residual error between the ETF phase shift and mean PDΣΔM output. For A_{ETF} = 1mV, g_M=1mA/V, C_{INT}=5.5pF, N=16, F_S=1.17MHz and $\Phi_{ETF} - \Phi_{0,1}$=1°, the integrated voltage error is 21.7mV. This means that 10mV comparator offset results in only 0.5° phase error for a coarse conversion. This is negligible compared to the coarse conversion's 11.25° resolution. A pre-amplifier can achieve this requirement within a small area. To save further area, the pre-amplifier also senses the common-mode of the gm-stage. The auxiliary branch formed by M5 and M6 detects the input common mode of the pre-amplifier with respect to the common-mode reference voltage (V_{CMREF}) and feeds the error back to the gm-stage through V_{CMFB}. In the end, the comparator consumes less than 10μA from a 1.8V supply and its area and current consumption overhead is small compared to the Gm-stage.

4.3.3. Digital Heater Drive Logic

Timing errors in the phase references Φ_0 and Φ_1 will directly lead to temperature errors. To minimize these, both F_{DRIVE} and F_{DEM} signal are synchronized by the high-frequency clock F_{SYNC} using D flip-flops (FF). This is shown in Figure 4.9, together with the H bridge (M1-4) used to drive the ETF heater (R_{HEAT}). Tapered inverters are used to drive the relative large switches M1-4, which have an on resistance of only 50 Ω each. When F_{DRIVE} is forced high, either M4-M1 or M2-M3 are on depending on the HDI signal and current flows through R_{HEAT}. When F_{DRIVE} is low, all switches are off, and the voltage on the resistor is floating between VDD and GND.

Having a floating voltage on the resistor minimizes the timing delay of the switches because it ensures that both PMOS and NMOS switches simultaneously

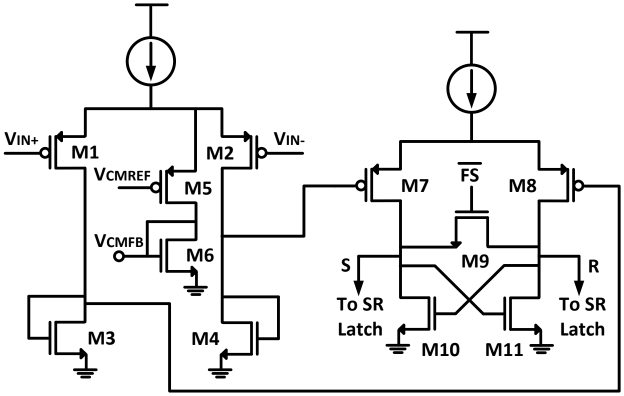


Figure 4.8: The implementation of the dynamic comparator

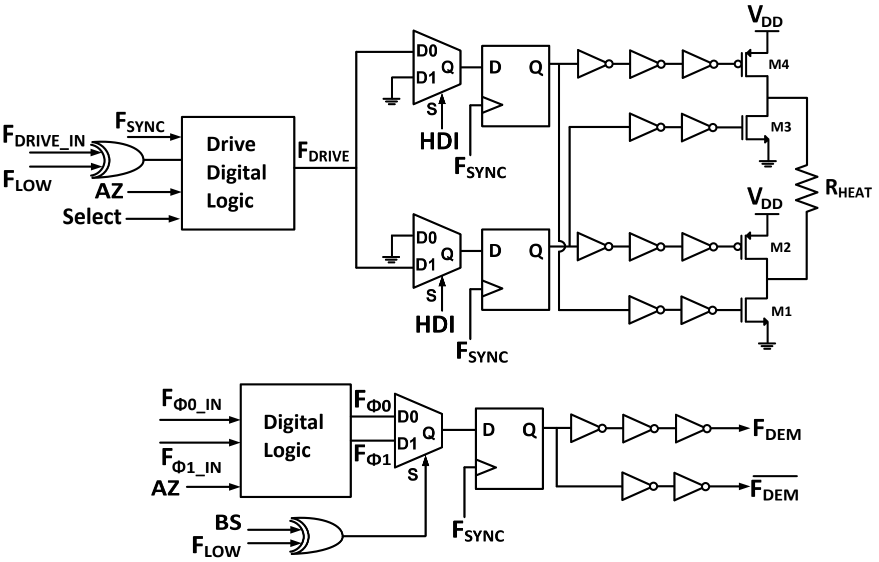


Figure 4.9: The implementation of the heater drive logic, which drives both the ETF heater (RHEAT) and the chopper drive signals going to the gm-demodulator.

conduct current at the very start of the heat pulse. This would not happen if the voltage on R_{HEAT} had been forced to ground or VDD between heat pulses. The exact voltage at which R_{HEAT} floats is of no importance since the signal of interest is encoded in the thermal domain. From simulations, the additional error in the PDΣΔM due to the delay of the digital logic is estimated to be roughly 0.4° at a drive frequency of 1.17 MHz (0.9 ns delay in time), which translates to 0.5° C inaccuracy error in temperature.

While the gm-stage is auto-zeroed, F_{DRIVE} is set to a relatively high frequency ($F_{SYNC}/2$). At this frequency, the ETF's AC output is quite small, while the same self-heating-induced DC offset is present as in normal operation [3]. To avoid delay errors, synchronizing flip-flops are also used to generate F_{DEM}. The expected mismatch between the delay of the synchronizing FFs and the tapered inverters causes a temperature error of about 0.5 °C at F_{DRIVE} = 1.17 MHz. In addition, the electrical phase shift due to the ETF's electrical parasitics and the limited BW of the gm-stage are estimated to contribute a worst-case error of 0.4 °C. Thus, it is expected that the sensor's accuracy will be dominated by the expected inaccuracy of the ETF itself (> 1.5 °C).

4.3.4. Digital Phase Reference Generator

Accurate generation of phases Φ_0 and Φ_1 are an important part of the design and can be easily done by using the synchronization clock F_{SYNC}. Figure 4.10 shows the implementation of the phase reference generation. F_{SYNC} is divided into 64 via a ripple counter, to generate F_{DRIVE}. Then, a 90° phase shift is applied for phase demodulation, and the resultant clock is given to two 16-element unary phase DACs. Each DAC has a step of 5.625° and can span 0 to 84.375°. The DAC control words REF0 and REF1 are loaded into the chip via an on-chip shift register. Outputs of the DACs are assigned to Φ_0 and Φ_1 phases and distributed to all the sensors on the chip.

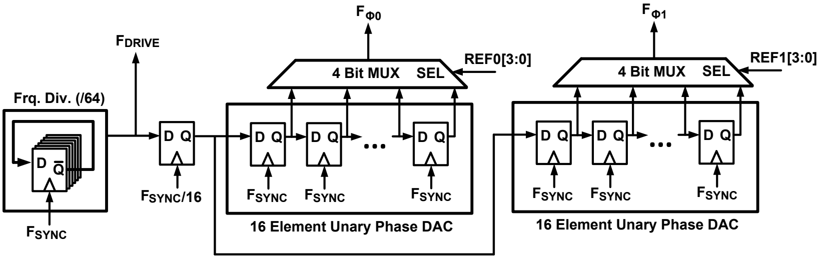


Figure 4.10: Implementation of the digital phase reference generator

As described before, the DACs are set to their maximum phase span for the coarse conversion. During the fine conversion, REF0 and REF1 are selected to have a difference of 2 elements or 11.25°. Therefore, the shift registers controlling the DACs are loaded twice per conversion. During measurements, external logic (on an FPGA) is used to program the shift registers before and during a conversion.

4.4. On-Chip Integration of an Array of TD Sensors

Using a compact temperature sensor opens up interesting architectural possibilities. An array of temperature sensors can be implemented on the same die. In this work, an array of 12 sensors has been realized.

Fig. 4.11 shows the block diagram of the proposed temperature sensing system.

It consists of an array of TD sensors distributed over the die, as well as shared circuitry for the generation of phase references and bias currents (not shown). Each sensor consists of an ETF and a PDΣΔM, which digitizes the ETF's temperature-dependent phase shift with respect to the phase references. As shown before in Figure 4.1, a single TD sensor block requires at least three minimum inputs: F_{DRIVE} as the ETF drive signal, as well as $F_{\Phi 0}$ and $F_{\Phi 1}$ to as act the phase reference signals. An external 75-MHz clock (F_{SYNC}) is used to generate these drive and phase reference signals within a single digital block. The phase shift of the references can be independently adjusted in 5.625° steps spanning a phase range from 0 to 84.375°.

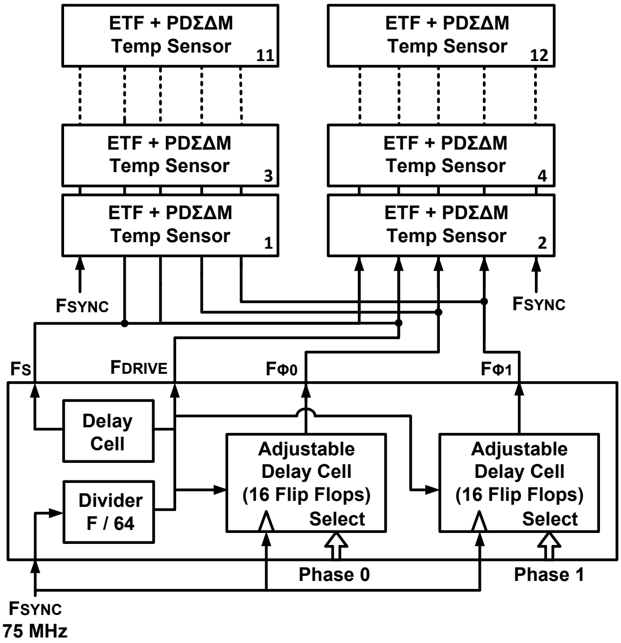


Figure 4.11: Block diagram of the system incorporating 12 temperature sensors and a shared phase-reference generator

The array consists of 6 sensors with n+ active heater ETFs and 6 sensors with n-well heater ETFs. Because of the poor noise performance of n-well heater ETFs, we will focus on ETFs with n+ active heaters.

4.5. Top Level Implementation

The proposed array of TD sensors was laid out and fabricated in SSMC 0.16 μm CMOS technology. The layout of a single TD sensor is shown in Figure 4.12 and the die photo is shown in Figure 4.13. The drawn dimensions of one sensor are 60 μm by 135 μm, which results in a silicon area of ~ 8000 μm^2.

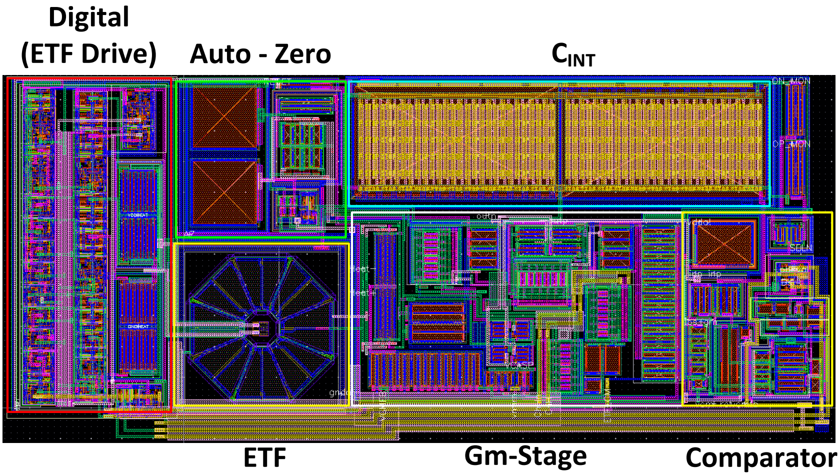


Figure 4.12: Layout of an individual sensor

The main purpose of this TD sensor is to detect fast on-chip thermal transients. To demonstrate this, a number of on-chip test heaters were realized, as shown in Fig. 4.13. The TD sensors are continuously monitored while these test heaters are turned on and off rapidly, which generates a transient thermal map of the chip.

4.6. Measurement Results

This section describes the measurement results of the prototype chips containing arrays of TD sensors. The chips were packaged in ceramic DIP24 packages, and tested from -40 to 125 °C in a controlled temperature environment. During measurements, it was found that the n-well heaters are 3x more resistive than predicted by its device model. Since the resistors are neat minimum size, this was attributed to pinch-off effects. This results in a low ETF signal amplitude andpoor resolution and accuracy. For this reason, the focus of the measurement results is on the ETFs with N+ active heaters.

4.6.1. Measurement Setup

To reduce measurement time, the prototype TD sensors were tested on a PCB board with slots for 4 chips. A Cyclone II FPGA interfaces with the chips and an off-board NI6537B DAQ card, which transfers the bit-stream of sensors directly to a PC-LABVIEW environment. There, the bit-stream is decimated, and the results are stored for further processing. The FPGA also handles all the I/O digital clock generation for the timing of the chips; as well as hosting the algorithm used for the two-step conversion. This approach gives flexibility in the implementation of the timing diagram shown previously in Fig. 4.4. A 50ppm accuracy MEMS clock source, operating at 75 MHz, is the reference precision clock used for the timing on

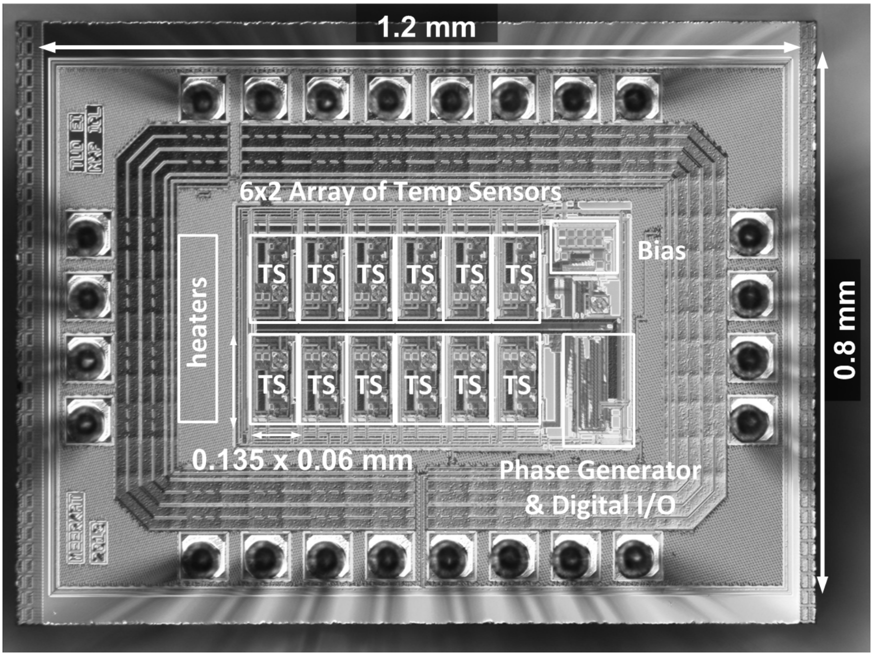


Figure 4.13: Die photo of the chip

board [5]. Figure 4.14 shows the block diagram of the measurement setup.

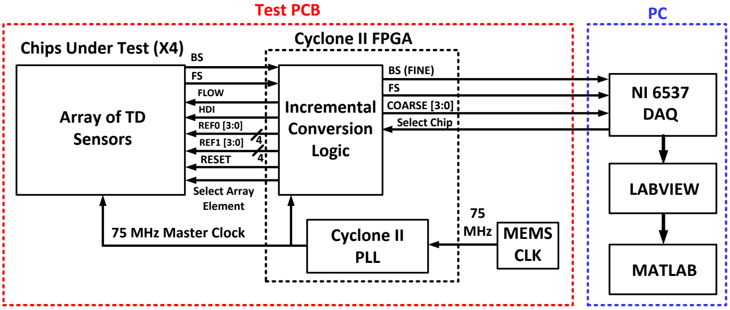


Figure 4.14: Block diagram of the measurement setup used for testing the prototype chips

As described before in Fig. 4.4, a conversion is divided into four sub-sections and takes 1130 clock cycles. When F_{SYNC} = 75 MHz, F_{DRIVE} is 75/64 = 1.172 MHz and thus one conversion takes place in 964 μs, or roughly 1 ms. The FPGA first runs a coarse conversion for 64 cycles and calculates a 4-bit value that approximates the ETF phase. Inside the FPGA, decimation of the 64 bits into a 4-bit value is done by a 6-bit counter whose two LSBs are ignored. A 6-bit look-up table is used to correct for the significant systematic non-linearity inherent in a PDΣΔM operation with a large phase span [6]. After this correction, REF0 and REF1 are selected as ± 1 values of the 4-bit coarse result, and the result is loaded to the chips to start a fine conversion. Then, the next 1024 received bits are sent to the PC to be decimated, along with the 4-bit coarse value.

4.6.2. ETF Phase-to-Temperature Curve

The phase-to-temperature behavior of the ETFs was determined by measuring their phase shift at 9 temperature points between -40 to 125 °C. A calibrated pt100 temperature sensor was used as a reference, and embedded in a large aluminum block in good thermal contact with the chips. In order to average out thermal noise, 100 measurements were obtained per temperature point per sensor. The mean of these plots, describing the average phase vs. temperature characteristic of each ETF, is defined as that ETF's master curve.

Fig. 4.15 shows the master curve of ETFs with n+ active heaters, which is obtained by averaging the output of 16 chips (96 sensors) from -40 to 125 °C. The master curve is a fifth-order polynomial used to convert the output of the PDΣΔM into absolute temperature. As expected, the master curve can be well approximated by the T^x law typical of TD sensors, where x ~ 0.9. TIn this design, it was found x ~ 0.83, which demonstrates that the ETF follows the predicted physical behavior.

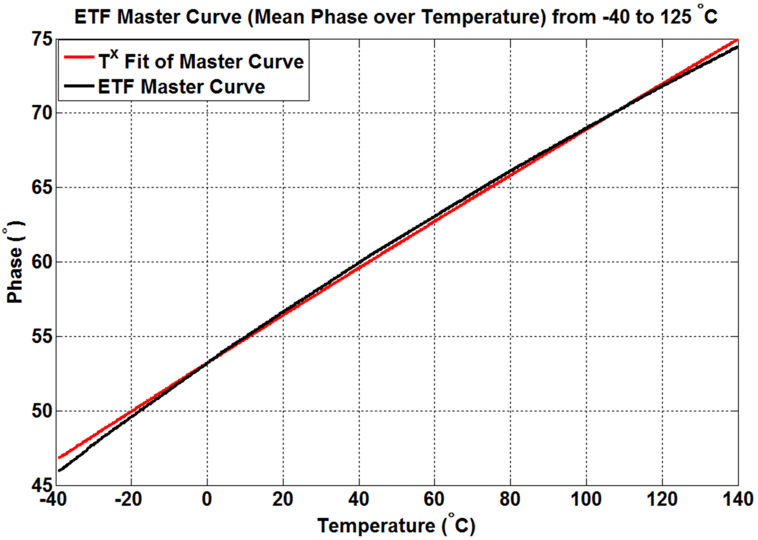


Figure 4.15: Master curve of n+ active ETFs from -40 to 125 °C, along with a T^x fit

4.6.3. Resolution

Fig. 4.16 shows the FFT of the bit-stream during a fine conversion. The thermal noise floor corresponds to a resolution of 0.21 °C (rms) in a conversion time of 1 ms, which was measured by obtaining the standard deviation of 10000 conversions at room temperature. The tones at 1 kHz and 2 kHz are due to system-level chopping and HDI. Both tones are strongly suppressed when a simple counter is used as the decimation filter for the incremental ADC. The power consumption of the ETF is 2.5 mW, while readout consumes 0.6 mW from the 1.8V supply.

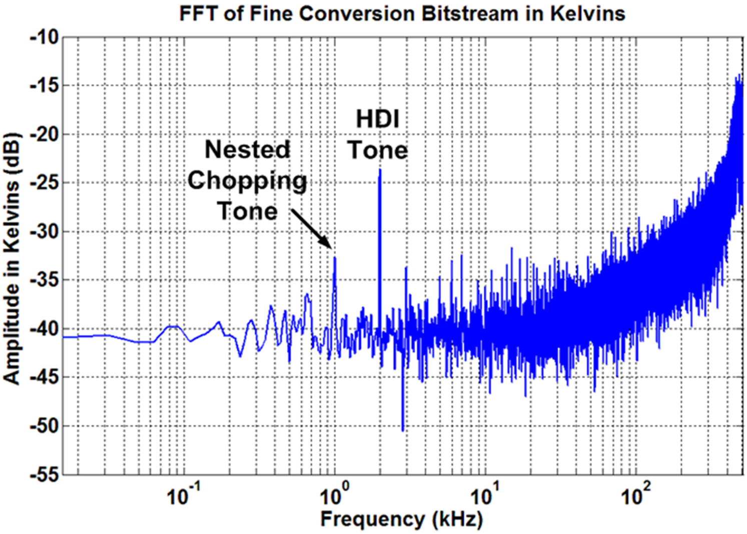


Figure 4.16: FFT of the PDΣΔM's bitstream during fine conversion for n+ active ETFs

4.6.4. Inaccuracy

As shown in Fig. 4.17, the untrimmed inaccuracy of 16 chips (96 sensors) is ±2.4 °C (3σ) from -40 to 125 °C for sensors with n+ active heater ETFs. However, the relative inaccuracy of sensors on the same die was found to be less than ±1.5 °C (3σ), which is shown in Fig. 4.18. The reduced spread exhibited by sensors from the same die allows a simplified calibration scheme, i.e. trimming only one sensor per die and using the same calibration parameters for all sensors. Such a scheme would be much faster, and thus cheaper than individual calibration, especially for a large number of sensors per die.

As shown in Fig. 4.19, the sensor's absolute inaccuracy drops to ±0.65 °C (3σ) after a 1-point digital offset trim at 70 °C, which is the typical trimming temperature for microprocessors [7]. In addition, the spread due to the self-heating of the ETFs is estimated to rbe about ±0.5 °C (3σ), which is part of the ±2.4 °C value reported above. This is approximately 20% of the total self-heating of the ETF (~ 2 °C) and occurs due to the 20% spread in the absolute value of the ETF's heater resistance.

4.6.5. Thermal Transient Response and Thermal Interference

As mentioned before, the test chip also includes an on-chip test heater. The response of 6 sensors to a 0.4-W temperature step generated by the on-chip heaters is shown in Fig. 4.20. The nearest sensor observes a transient with a slope of

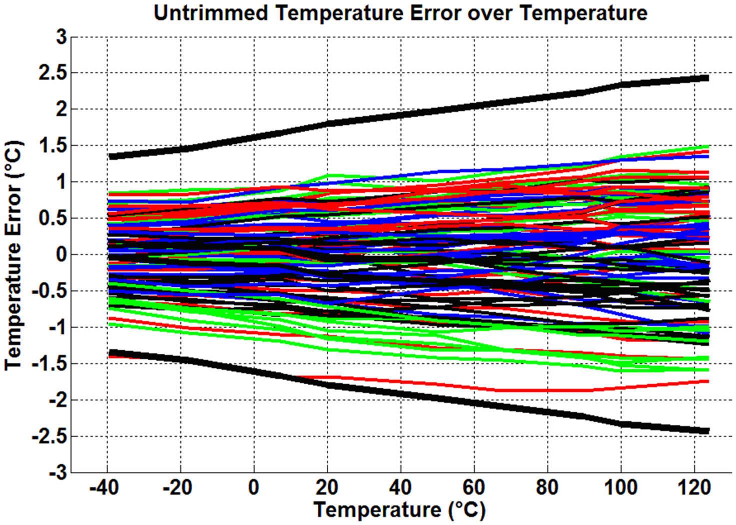


Figure 4.17: Measured sensor-to-sensor temperature errors for 16 chips (96 sensors) without any temperature trimming. Black bold lines indicate 3σ limits.

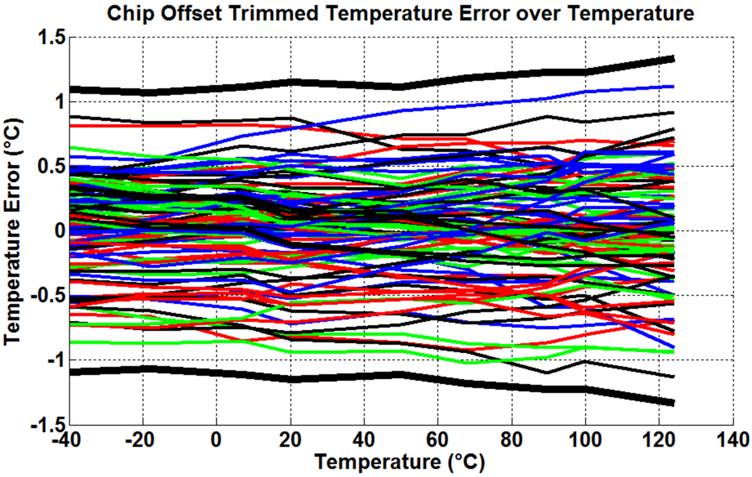


Figure 4.18: Measured sensor-to-sensor temperature errors for 16 chips (96 sensors) after trimming only one sensor per chip at 70 °C and using the same trim result for all the sensors on the same chip. Black bold lines indicate 3σ limits.

1 °C/ms, while the other sensors observe progressively smaller transients as expected. This validates the ability of the sensor to detect rapid thermal transients.

To investigate the ETF's sensitivity to thermal interference, the on-chip heaters were driven by a 0.4-W pseudo-random sequence derived from F_{DRIVE}. Apart from a baseline shift due to the increase in die temperature, no extra noise was observed

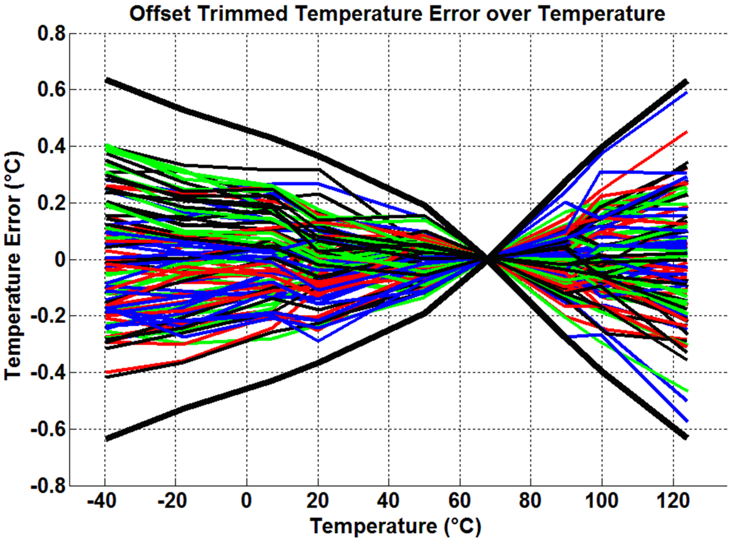


Figure 4.19: Measured sensor-to-sensor temperature errors for 16 chips (96 sensors) after single-point trimming at 70 °C. Black bold lines indicate 3σ limits.

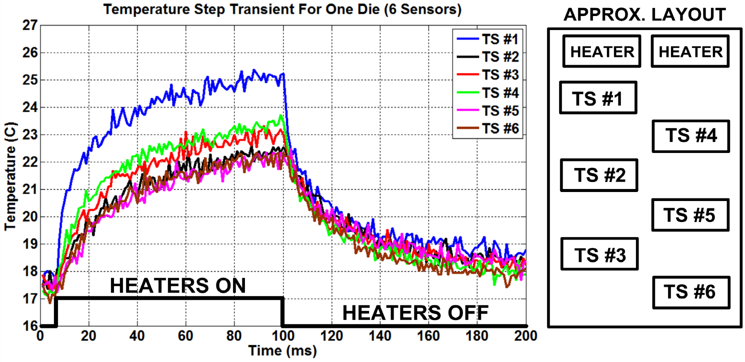


Figure 4.20: Step response of six temperature sensors within one die when turning the resistive test heaters on and off. Locations of the sensors with respect to the heaters are also shown.

in the sensors' outputs, which demonstrates that such interference is effectively filtered out by the substrate's thermal inertia.

4.6.6. Linearity and Ramp Measurements

To characterize the non-linearity of the PDΣΔM, a ramped temperature measurement was done from -40 to 125 °C. Figure 4.21 shows the statistical averages obtained from a 50 mK/sample ramp. The results were averaged into 0.5 °C segments to reduce noise. The bold black lines indicate the estimated 3σ accuracy limits including thermal noise and INL errors, while the red line indicates the mean

INL error due to the gain error of the PDΣΔM. It can be seen that a deterministic error occurs near the transitions of the coarse conversion. This error is constrained to within ±0.2 °C and is roughly equal to the sensor's noise-limited resolution. The cause of these errors was traced to the limited gain of the gm stage used in the PDΣΔM, causing a gain error in the phase-to-digital conversion. Simulation results of the 1st-order PDΣΔM predict an INL error of ~ 0.4 °C (peak-to-peak) for a 5.5pF integration capacitor and a 1 mA/V gm-stage with the targeted DC gain of 80 dB.

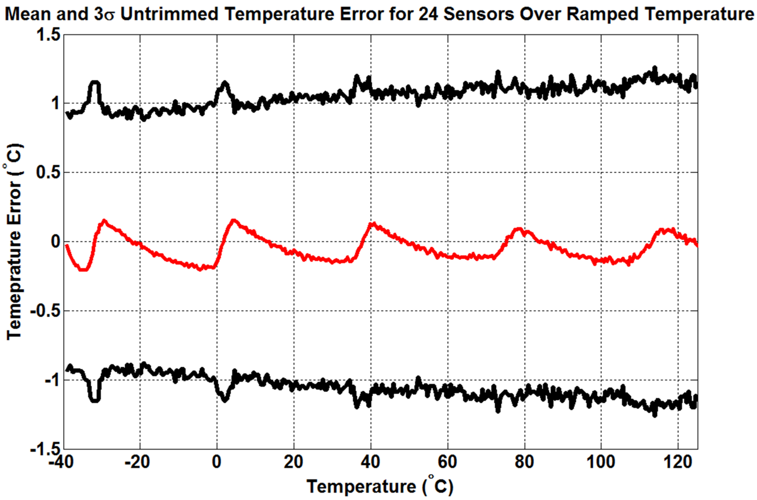


Figure 4.21: The average temperature error and the 3σ inaccuracy of 24 sensors with respect to the general master curve for an oven ramp from -40 to 125 °C

The INL errors are caused by the fact that a ΣΔ modulator with a leaky integrator (limited DC gain) will exhibit gain error [8]. For a normalized input value of x, this error is (1-p)*x [8], where the integrator leakage for a gm-C integrator is [2]:

$$p = e^{-1/(F_S R_O C_{INT})}$$

Here, F_S is the sampling frequency, R_O is the output impedance of the gm-stage and C_{INT} is the integration capacitance. In a two-step converter, the INL errors occur at the transitions of the coarse converter. The output of the coarse conversion will then toggle by one LSB, leading to normalized fine conversion outputs of either x or (x-1). Thus, the total normalized error is then,

$$(1-p)*x-(1-p)*(x-1)=(1-p)$$

A graphical representation of this error is shown in Fig. 4.22. Here, an integer change in the x-axis (normalized input) corresponds to one LSB of the coarse conversion. Gain error in the fine conversion steps translates into large DNL and INL errors.

For an LSB step is 5.625° and a DC gain of 80 dB (8 dB lower than this implementation) this non-ideal behavior leads to an INL error of ±0.25 °C, i.e. slightly

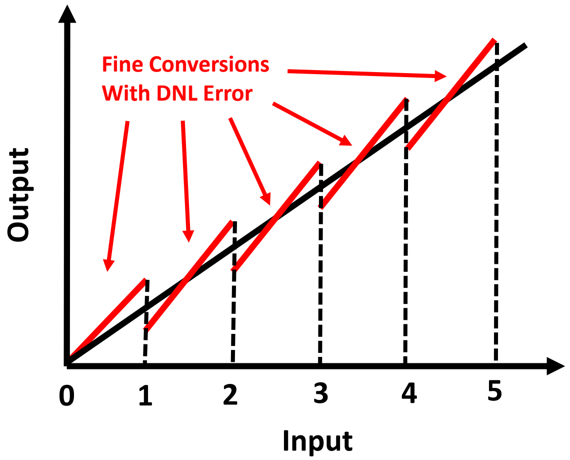


Figure 4.22: Graphical representation of the DNL/INL errors in a two-step converter, shown via in input to output plot of an ideal ADC.

larger than the measured error. This is not a problem, since it is of the same order as the ETF's thermal noise and is well below its inaccuracy. However, it may pose a challenge for future designs in nanometer CMOS, in which it may be more problematic to achieve sufficiently high DC gain.

4.6.7. System Level Chopping and HDI

In order to test the efficacy of the proposed low-frequency chopping and HDI techniques, 4 chips (24 sensors) with an estimated 3σ of 1.2 °C were tested from -40 to 125 °C by disabling HDI and low-frequency chopping during the tests. Figure 4.23 shows the untrimmed inaccuracy of the 24 sensors from -40 to 125 °C in 4 possible modes of operation: no HDI/no low-frequency chopping, no HDI with low-frequency chopping, with HDI but no low-frequency chopping, and both HDI and low-frequency chopping, respectively denoted as modes 1-4. It can be observed that the effect of HDI is minimal, since it only slightly reduces the spread at cold temperatures, thus leading to the conclusion that the effect of crosstalk is negligible. However, the absence of low-frequency chopping has a dramatic effect and increases the estimated 3σ inaccuracy of these 24 sensors to 2.3 °C.

4.7. Conclusion

A compact TD sensor in 160-nm CMOS has been described, and techniques which allow the sensor to be implemented in a compact area have been discussed. The sensor's area, speed, and resolution satisfy typical specifications for SoC thermal monitoring, while its untrimmed inaccuracy is the lowest reported for temperature sensors targeting this application. Since the performance (area, accuracy, power, speed) of TD sensors is expected to improve with process scaling, these results

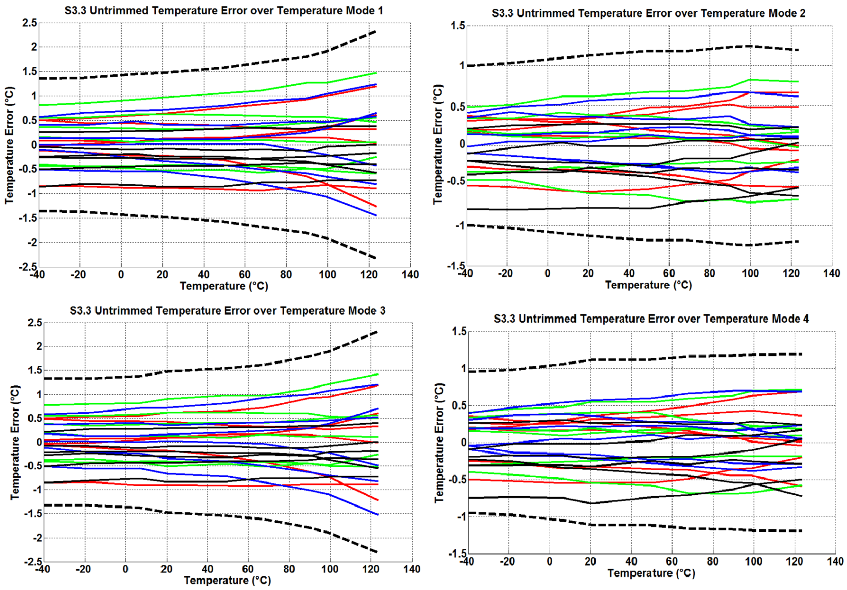


Figure 4.23: Inaccuracy of 24 sensors with 4 modes of enabling HDI and low frequency chopping. Black bold lines indicate estimated 3σ limits. The 4 modes are respectively: no HDI or low frequency chopping, only low frequency chopping, only HDI and both enabled.

demonstrate that they are a good choice for hot-spot monitoring in microprocessors and other systems-on-chip. With respect to previous TD sensors [3] [2], the proposed sensor is 22x smaller and 1000x faster. Furthermore, its resolution Figure of Merit (FOM) is also 55x better, improving from 7498 nJK^2 in [3] to 137 nJK^2 in this work. These results show that the proposed design techniques have been successful in realizing TD sensors for thermal monitoring. The following chapter describes a further implementation in standard 40nm CMOS and demonstrates the benefits of scaling for TD-based temperature sensors.

References

[1] C. van Vroonhoven, D. D'Aquino, and K. Makinwa, "A ±0.4 °C (3σ) -70 to 200°C time-domain temperature sensor based on heat diffusion in Si and SiO2," in *Digest of Technical Papers ISSCC*, Feb 2012, pp. 204–206.

[2] S. Kashmiri, S. Xia, and K. Makinwa, "A Temperature-to-Digital Converter Based on an Optimized Electrothermal Filter," *IEEE Journal ofSolid-State Circuits*, vol. 44, no. 7, pp. 2026–2035, July 2009.

[3] C. van Vroonhoven, D. D'Aquino, and K. Makinwa, "A thermal-diffusivity-based temperature sensor with an untrimmed inaccuracy of ±0.2 °C (3σ) from -55°C to 125°C," in *Digest of Technical Papers ISSCC*, Feb 2010, pp. 314–315.

[4] K. Makinwa and M. Snoeij, "A CMOS Temperature-to-Frequency Converter With an Inaccuracy of Less Than 0.5°C (3σ) From - 40°C to 105°C," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 12, pp. 2992–2997, Dec 2006.

[5] *Low Jitter Pin Configuration CMOS Output 3.2 x 2.5 x 0.85 mm Ultra Miniature Pure Silicon Clock Oscillator*. [Online]. Available: http://www.abracon.com/Oscillators/ASEMCC.pdf

[6] C. van Vroonhoven and K. Makinwa, "Thermal diffusivity sensors for wide-range temperature sensing," in *IEEE Sensors*, Oct 2008, pp. 764–767.

[7] E. Rotem, J. Hermerding, A. Cohen, and H. Cain, "Temperature measurement in the Intel(R) CoreTM Duo Processor," in *THERMINIC*, 2006.

[8] O. Feely and L. Chua, "The effect of integrator leak in Σ-Δ modulation," *Circuits and Systems, IEEE Transactions on*, vol. 38, no. 11, pp. 1293–1305, Nov 1991.

5

Compact Smart TD-Based Temperature Sensors in 40nm CMOS

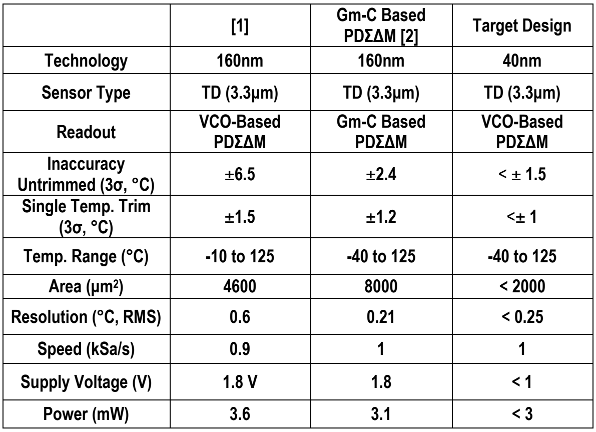
This chapter presents the design and measurement results of an array of 1650 μm^2 1 kSa/s thermal-diffusivity based temperature sensors in 40nm standard CMOS. They achieve inaccuracies down to ±1.45 °C (3σ) from -40 to 125 °C with no trimming and ±0.75 °C (3σ) after a single-point temperature trim. They also achieve resolution of 0.36 °C for a power consumption of 2.5 mW. These results demonstrate the feasibility of TD sensors in nanometer CMOS.

5.1. Introduction

As shown in Chapter 4, thermal-diffusivity (TD) based temperature sensors have been successfully implemented for thermal management applications in 0.16 μm CMOS. Scaling of this design to a more modern CMOS process is the next step, since modern SoCs and CPUs are exclusively implemented in such processes. Therefore, the objective of this chapter is to demonstrate a TD sensor designed in 40nm CMOS, which benefits from scaling down from 160nm.

In Chapter 3, two different phase domain ΣΔ modulator (PDΣΔM) architectures were discussed in terms of their compatibility with technology scaling. Gm-C PDΣΔMs employ mainly analog circuit blocks, while the more digital-friendly VCO-based PDΣΔMs suffer from additional quantization noise. Table 5.1 compares the performance of a VCO-based TD sensor with the Gm-C based design presented in Chapter 4.

Table 5.1: Comparison table showing previous state-of-the-art TD sensors and design targets in 40nm



	[1]	Gm-C Based PDΣΔM [2]	Target Design
Technology	160nm	160nm	40nm
Sensor Type	TD (3.3µm)	TD (3.3µm)	TD (3.3µm)
Readout	VCO-Based PDΣΔM	Gm-C Based PDΣΔM	VCO-Based PDΣΔM
Inaccuracy Untrimmed (3σ, °C)	±6.5	±2.4	< ± 1.5
Single Temp. Trim (3σ, °C)	±1.5	±1.2	<± 1
Temp. Range (°C)	-10 to 125	-40 to 125	-40 to 125
Area (µm²)	4600	8000	< 2000
Resolution (°C, RMS)	0.6	0.21	< 0.25
Speed (kSa/s)	0.9	1	1
Supply Voltage (V)	1.8 V	1.8	< 1
Power (mW)	3.6	3.1	< 3

From the table, it can be seen that the VCO-based design has significantly worse performance: 2.5x worse untrimmed inaccuracy, 3x worse resolution and somewhat higher power consumption. The main reasons for this are as follows [1]:

1. As shown in Fig. 5.1, the front-end VCO is outside the ΣΔ loop. Hence its delay adds to the ETF phase shift, and increases its spread.

2. As explained in section 3.6.1, the use of a digital counter as the modulator's integrator imposes rounding (or quantization) at its input and introduces an additional noise source.

The last column in Table 5.1 describes the scaling targets of a design in 40nm. Due to process scaling from 160 to 40 nm, the area is expected to reduce by a factor of roughly 16, and the voltage supply will drop to approximately 1 V. However, as is common for analog circuits, a complete scaling factor of 16x is difficult to achieve, and a more modest 4x scaling factor is assumed.

The power consumption and resolution of the design in [2] are adequate for the application and resolution of [2] is limited by time-domain quantization noise. Therefore, these specifications are not targeted for improvement. The accuracy target was chosen as 1 °C (3σ), based on the requirements of the intended thermal management application [3].

Therefore, the main goal of designing this TD sensor is to exploit the small area of the VCO-based PDΣΔM, while eliminating its additional noise and accuracy penalties. Furthermore, a secondary goal is to push the VCO-based PDΣΔM design to achieve the resolution and accuracy performance of the gm-C readout in 0.16μm technology while benefiting from smaller area of the VCO-based architecture in 40nm.

5.2. System Level Overview

The block diagram of the proposed VCO-based architecture is shown in Fig. 5.1, and its operation is summarized as follows. An ETF is driven at a drive frequency F_{DRIVE} and generates a small voltage signal V_{ETF} (also at the same drive frequency) at a phase shift of Φ_{ETF} from F_{DRIVE}. A front-end that acts as a VCO converts V_{ETF} into a frequency signal F_{VCO}, which still keeps the phase shift Φ_{ETF} after frequency modulation. A digital PDΣΔM consisting of an up/down counter, phase DAC and digital quantizer is used for digitizing Φ_{ETF} over F_{VCO}. A more detailed description of the system is given in section 3.6. Its main advantage is that it mainly uses digital components which scale with technology.

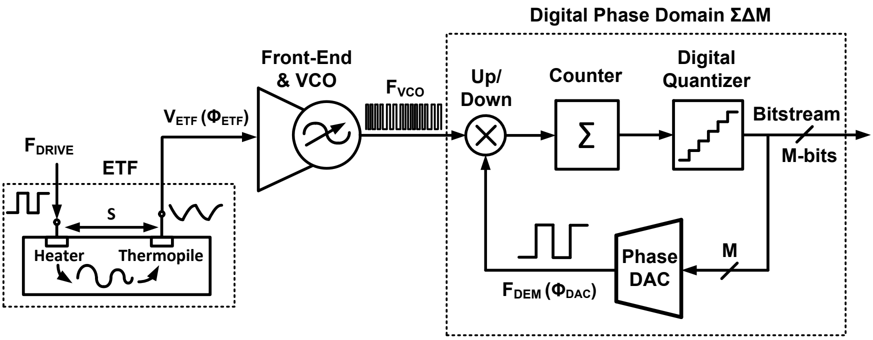


Figure 5.1: VCO based PDΣΔM architecture

As discussed in Chapter 3, design of a VCO-based PDΣΔM starts by determining three important system parameters: S (counter length), M (the number of quantizer bits) and K_{VCO} (front-end VCO gain in MHz/mV). K_{VCO} is an important parameter that determines quantization noise, wrap-around and stability (sections 3.6.1 and

3.6.2).

In this design, both s3.3 and s2 polygon ETFs will be used. These ETFs are described in detail in sections 2.7.1 and 2.7.2. The s3.3 ETF was previously implemented in a mature 160nm CMOS process (see chapter 4), and so reimplementing it in 40nm CMOS allows us to directly observe the effect of process scaling. The s2 ETF may be seen as a scaled version of s3.3 ETF.

The PDΣΔM should be designed to match the expected ETF performance. First, we consider the case of the s3.3 ETF with F_{DRIVE} = 1.17 MHz, thermopile resistance of ~8 kΩs and heater power of 2 mW. For these values, the ETF signal is a filtered square-wave with an amplitude of ~1.3 mVpp and a phase resolution of 35 m° in a bandwidth of 500 Hz. We will accept 30% more thermal noise from the front-end, which increases the total phase resolution to 47 m°. This also sets the target phase resolution of the front-end amplifier to 19 m° within 500 Hz. By choosing F_S=1.17 MHz, an oversampling ratio (OSR) > 1024 is obtained. The phase range of the readout is designed to span 11.25° to 90°, derived from the ETF characteristics over temperature, where we will define this phase range (78.75°) as Δ.

In order not to significantly reduce sensor resolution, we need to make K_{VCO} large enough to suppress the counter's quantization noise ($\sigma_{P,^\circ}$), as discussed in section 3.6.1. We choose a target $\sigma_{(P,^\circ)}$=23 m° [equation 3.26], i.e. roughly half of the signal noise, which results in a 12% SNR degradation. From equation 3.26, K_{VCO} is then found to be 180 MHz/mV. In practice, since K_{VCO} will spread over corners, K_{VCO} was chosen to be 200 MHz/mV for the typical case, and 160 MHz/mV in the worst-case. Since we want to observe the worst cases for both resolution and wrap-around, we assume K_{VCO} = 160 MHz/mV for quantization-noise calculation and 200 MHz/mV for wrap-around estimation. K_{VCO} = 200 MHz/mV results in a total computed RMS resolution of 54 m° [47 m° due to sensor noise; 26 m° due to counter quantization noise].

The next step is determining S from equation 3.32 and 3.34. Δ was defined as 78.75°, resulting in S > 6.12 from equation 3.34. For equation 3.32, we need to fix M, or the number of ΣΔ modulator bits. M=3 was chosen as a good trade-off between phase DAC area and quantization noise suppression. The mixed-signal CppSim [4] model of the architecture shown in Fig. **??** for K_{VCO} = 200 MHz/mV and an ETF phase (Φ_{ETF})=42° was used to obtain a histogram of the counter output swing (Fig. 5.2). The peak-to-peak swing is 23 count values for 8192 samples. According to equation 3.32, 2^{S-M} > 23 to avoid wrap-around, which implies S ≥ 8. This satisfies the requirement from equation 3.34 as well.

5.3. Foreground Phase Calibration

One of the disadvantages of a VCO-based PD ΣΔM is the additional phase shift introduced by the open-loop VCO, which consists of a Gm-stage followed by a CCO. Power and area considerations mean that these stages cannot be made arbitrarily fast, and therefore we are left with an accuracy penalty.

One way of eliminating this error is to use a foreground calibration technique. As such, the readout's phase error can be actively monitored and cancelled in the digital back-end. This so-called phase calibration method can also provide statistics

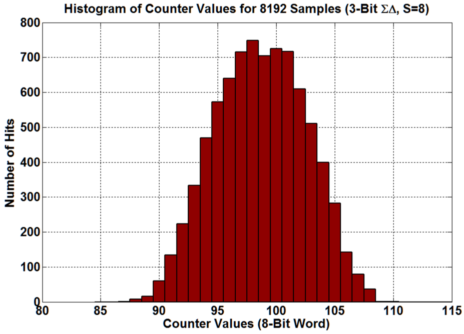


Figure 5.2: Histogram of counter values for 8192 samples for 3-bit ΣΔ, 8-bit counter, phase range of 78.75°.

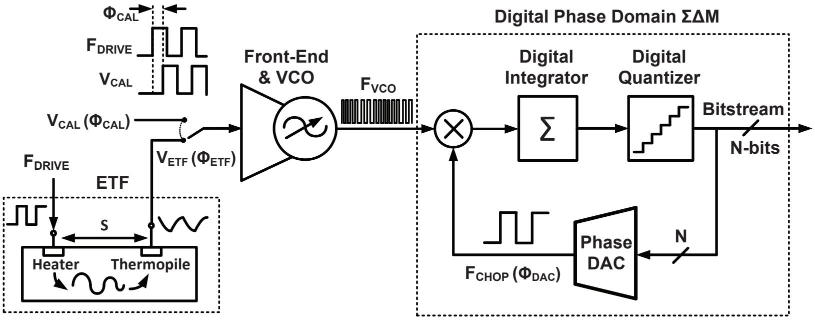


Figure 5.3: System level block diagram of the VCO-Based PDΣ Δ M with Phase Calibration

of the readout phase delay and its spread, making it a diagnostic test tool as well.

In order to rule out any circuit-related non-idealities and non-linearity errors, the input signal to the readout should behave similarly to the ETF signal during this foreground calibration. One possibility is to inject a small voltage pulse to the input of the gm-stage, at a frequency F_{DRIVE} and a known phase shift. We will call this signal V_{CAL} and its phase shift the reference calibration phase or Φ_{CAL}. Figure 5.3 shows a simplified block diagram of a PDΣΔM with phase calibration. Via the CAL signal, the sensor can be made to operate in either a normal or a calibration mode.

This foreground calibration technique gives flexibility during measurement and characterization of the sensors by allowing continuous or one-time calibration of the readout error. Continuous calibration can eliminate temperature-dependent readout errors, but increases conversion time. One-time calibration can eliminate static readout errors (such as phase offset), but not temperature or drift related errors. In return, it is simpler and less time-consuming to perform. The reference phase for calibration was set to 22.5°, but it was chosen arbitrarily as any reference value is suitable as long as it is within the 78.75° phase range.

Φ_{CAL} can be readily obtained from the phase DAC used as part of the multi-bit feedback. A switched current source, denoted as ICAL, is used to convert this digital signal into V_{CAL} over the ETF's thermopile resistors. Therefore, only minimal circuit area is needed to implement phase calibration efficiently. For more flexibility, ICAL has been designed as a programmable current DAC to observe how the phase calibration results change with respect to non-ideal circuit behavior. The circuit-level implementation of the scheme will be described in the next section. The final system, including phase calibration, is shown in Fig. 5.4.

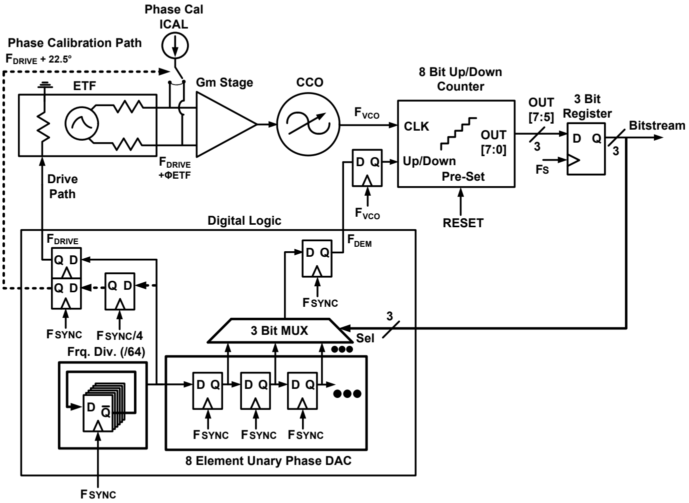


Figure 5.4: Detailed system-level block diagram of the implemented ETF + VCO based PDΣ Δ M readout

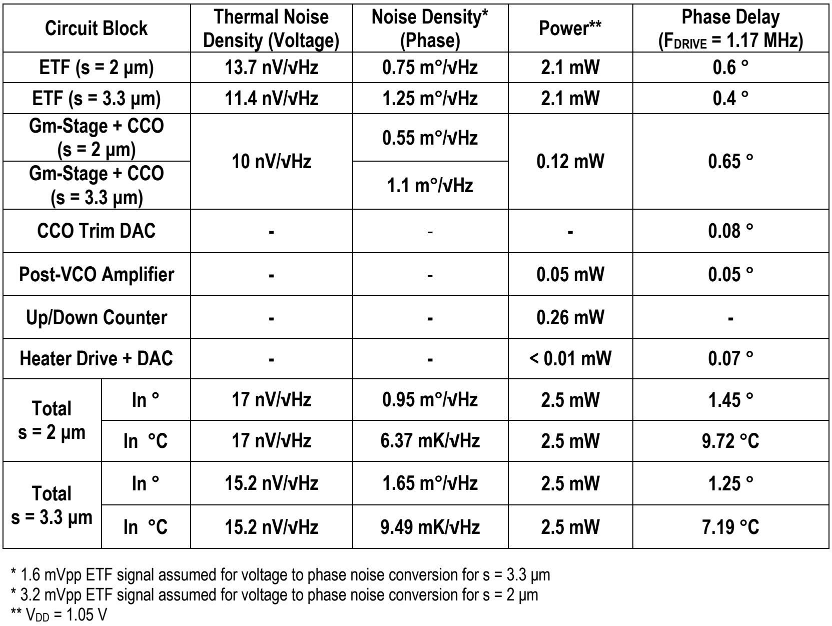
5.4. Circuit Design

Two performance specifications dominate the design of the readout circuitry: noise and accuracy. The former is dominated by the thermal noise of the Gm-stage, and the time-domain quantization noise of the VCO-based PDΣΔM as shown in section 3.6.1. The latter is determined by the delay of the open-loop blocks in the design: Gm-stage, CCO, post-CCO amplifier, and digital heater drive logic. Phase calibration aims to monitor and eliminate most of these error sources, but the delay of the DAC used for calibration adds a small delay and inaccuracy as well. Table 5.2 shows the overview of the circuit blocks used in the design and their noise and delay performance budgets. The analysis is made with both the s2 and s3.3 ETFs. Assuming that 20% mismatch of the total delay corresponds to 3σ inaccuracy, the untrimmed inaccuracy of the readout (without phase calibration) is estimated to be 1.9 °C for the s2 ETF and 1.4 °C for the s3.3 ETF.

Due to reduction of voltage supply from 1.8V to 1V, ETF power reduces from 2.5 mW to 2.1 mW despite reducing the ETF heater resistance, and this results in a

16% reduction of the output signal. Therefore, resolution due to ETF and readout thermal noise is estimated to be 0.23 °C (RMS) for s = 3.3μm and 0.15 °C (RMS) for s = 2μm ETF. K_{VCO}(VCO gain) will be kept high at 200 MHz/mV to prevent excessive time-domain quantization noise. Still, the quantization noise causes the estimated resolution to degrade slightly, down to 0.26 °C (RMS) for the s3.3 ETF and 0.17 °C (RMS) for the s2 ETF.

Table 5.2: Overview of circuit blocks and noise/accuracy specifications of these blocks in 40nm



Circuit Block		Thermal Noise Density (Voltage)	Noise Density* (Phase)	Power**	Phase Delay (F_{DRIVE} = 1.17 MHz)
ETF (s = 2 µm)		13.7 nV/√Hz	0.75 m°/√Hz	2.1 mW	0.6 °
ETF (s = 3.3 µm)		11.4 nV/√Hz	1.25 m°/√Hz	2.1 mW	0.4 °
Gm-Stage + CCO (s = 2 µm)		10 nV/√Hz	0.55 m°/√Hz	0.12 mW	0.65 °
Gm-Stage + CCO (s = 3.3 µm)			1.1 m°/√Hz		
CCO Trim DAC		-	-	-	0.08 °
Post-VCO Amplifier		-	-	0.05 mW	0.05 °
Up/Down Counter		-	-	0.26 mW	-
Heater Drive + DAC		-	-	< 0.01 mW	0.07 °
Total s = 2 µm	In °	17 nV/√Hz	0.95 m°/√Hz	2.5 mW	1.45 °
	In °C	17 nV/√Hz	6.37 mK/√Hz	2.5 mW	9.72 °C
Total s = 3.3 µm	In °	15.2 nV/√Hz	1.65 m°/√Hz	2.5 mW	1.25 °
	In °C	15.2 nV/√Hz	9.49 mK/√Hz	2.5 mW	7.19 °C

* 1.6 mVpp ETF signal assumed for voltage to phase noise conversion for s = 3.3 µm

* 3.2 mVpp ETF signal assumed for voltage to phase noise conversion for s = 2 µm

** V_{DD} = 1.05 V

5.4.1. Gm-Stage

The Gm-stage needs to interface with the variable input impedance of the CCO, which is typically 10-100 kΩ as shown in section 5.4.3. This necessitates a high output impedance from the gm-stage, which means a cascoded or two-stage design. Since sub-1V operation is desirable, a two-stage single-ended amplifier architecture was preferred over a folded cascode OTA. The two-stage design also requires fewer transistors (8) compared to a standard folded cascode (11), so it is also slightly smaller in area. A circuit level schematic of the proposed two-stage amplifier is shown in Fig. 5.5.

The first stage, consisting of transistors M1-4 forms a single-ended differential amplifier that provides 21 dB gain over a bandwidth of 200 MHz with a current consumption of 100 μA. Its 10 nV/$\sqrt{(Hz)}$ input referred noise density at 1 MHz is mostly

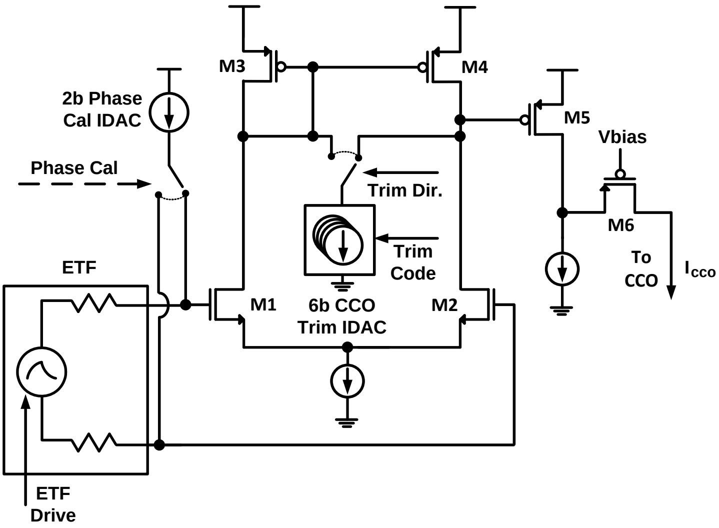


Figure 5.5: Schematic of the Gm-Stage

determined by the thermal noise of M1-2. The second stage improves the output impedance to ∼ 80 KΩ and increases the total transconductance of the stage to 1.5 mA/V. The cascode transistor M6 acts as a current buffer that interfaces the CCO's low input impedance with the gm's medium output impedance. It is dimensioned for an intrinsic gain of ∼15-20 at a small area, to avoid increasing the parasitic capacitance at its source. The Bode plot of the gm-stage's total transconductance (the current delivered to the CCO) over frequency is shown in Fig. 5.6. Its phase delay at the drive frequency (0.65° error) is added to that of the ETF, but is also detected and subtracted during the phase calibration mode.

The current consumed by the CCO is only several μA, and therefore the offset of the gm-stage can significantly alter the nominal frequency of the CCO, or even turn it off completely. Moreover, the frequency of the CCO will spread over PVT. To compensate for these errors, a 6-bit binary element IDAC, used in combination with a polarity switch, injects a current into the M1-2 pair. The IDAC exhibits a resolution of 0.5 μA, corresponding to an input-referred offset resolution of 0.5 mV and can cover a range of ±30 mV input offset. The external trimming logic used to program the IDAC and monitor the CCO frequency to converge into the desired frequency range is covered in section 5.6.2.

This trimming scheme also results in an unintended spread mechanism. Increasing the IDAC values to modify CCO frequency results in an increase of bias current through the pair M1-2, and hence increases the bandwidth of the Gm-stage. Therefore, the phase delay of the Gm-stage becomes a function of the trim IDAC setting (or CCO frequency). This can be seen in Figure 5.7, which shows the delay of the Gm-stage as the trim DAC value is changed. The simulated 205 ps delay

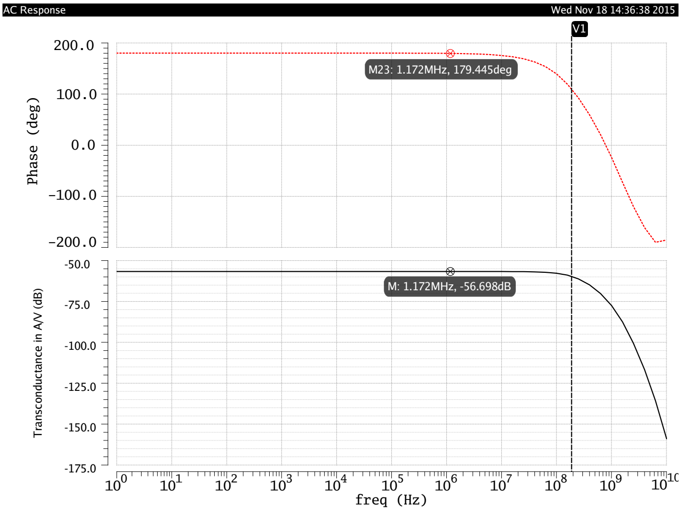


Figure 5.6: Bode plot showing the transconductance of the Gm-stage over frequency

difference results in 85 m° phase shift for 1.5 μA change in trimming current. As can be seen from the figure, the change in phase error is non-linear and follows an inverse relationship to the IDAC current.

5.4.2. Phase Calibration IDAC

The gm-stage also includes the 2-bit binary phase calibration IDAC (ICAL) which injects a current to the ETF thermopiles during calibration. When the calibration signal is enabled, the IDACs alternate between charging one terminal of the gm-stage, according to the phase calibration signal itself. Each calibration IDAC LSB injects 125 nA, which corresponds to 0.5 mVpp square wave signal over the ETF for an 8 kΩ differential resistance.

The delay of the IDAC is critical since it will only be present during phase calibration and will be subtracted from the ETF phase shift in the end. The total delay, for typical corner and room temperature is shown in Fig. 5.8. This 2.36 ns delay corresponds to 1 ° phase for a driving frequency of 1.17 MHz. However, 0.65° phase was due to the Gm-stage as was shown in Fig. 5.6. The remaining 0.35° error is due to phase calibration itself and corresponds to roughly 0.45 °C temperature inaccuracy. Therefore, we can expect phase calibration to detect and remove the 0.65° phase error due to Gm-stage, with a residual calibration error of 0.35°. In the temperature domain, this means a reduction of readout-related inaccuracy from 0.84 °C to 0.45 °C.

5.4.3. CCO

As in the GmC-based design, the CCO was implemented as a ring oscillator.

Fig. 5.9 shows an N-stage ring oscillator connected to an ideal current source I_{CCO}, with a voltage swing of V_{CCO} at a running frequency $F_{CCO} = 1/t_{CCO}$.

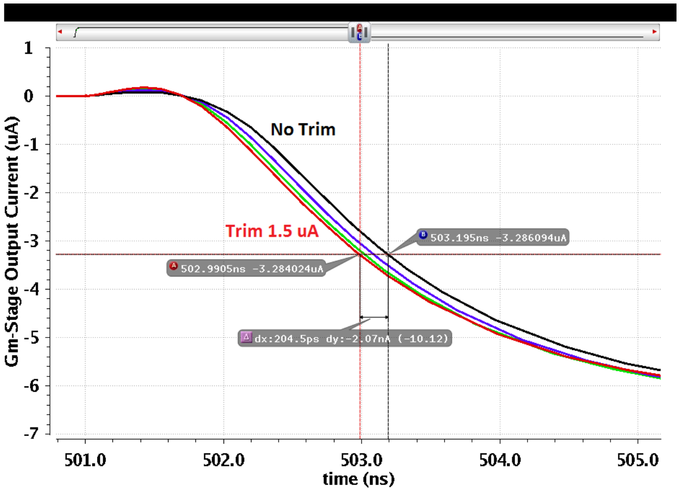


Figure 5.7: Transient simulation showing the varying delay of the Gm-stage to a pulse at its input, when the trim DAC value is changed. The black line corresponds to 500 MHz CCO frequency, while the red line corresponds to 800 MHz.

The first important specification for the CCO is its gain, denoted as K_{CCO}; which together with the transconductance of the gm-stage, determines the VCO gain K_{VCO}. As derived in section 5.2, K_{VCO} should be at least 160 MHz/mV in order to limit quantization noise.

Since the gm of the first stage (for noise considerations) is at least 1mA/V; the CCO gain should be at least 160 MHz/μA. Figure 5.10 shows the CCO frequency versus current for N=3,5 and 7 stages. As expected, the CCO frequency decreases as the number of stages are increased. Fig. 5.11 shows the CCO gain versus current, clearly demonstrating non-linear behavior. The gain of the 3-stage CCO is considerably higher than CCOs with more stages, and thus choosing N = 3 is favorable for power efficiency.

The non-linearity of the CCO gain is another concern since it can distort the input signal. While any systematic non-linearity would be captured as a modification of the non-linear phase-to-temperature curve associated with the ETF; the spread of the non-linearity due to process and temperature variations would appear as inaccuracy.

In order to study the non-linearity of the ring-inverter based CCO, let us first derive the oscillation frequency of the CCO. From figure 5.9, the oscillation frequency of such a structure can be found by considering that the mean current over the parasitic capacitors C must be zero. Therefore, assuming I_{CCO} is a current source with a nominal value of I_{NOM} and I_{IN} is the mean input current, or $I_{CCO} = I_{IN} + I_{NOM}$:

$$F_{CCO} = \frac{I_{IN} + I_{NOM}}{NV_{CCO}C} \tag{5.1}$$

If V_{CCO} or C are non-linear functions of I_{CCO} or F_{CCO}, as is the case in practice,

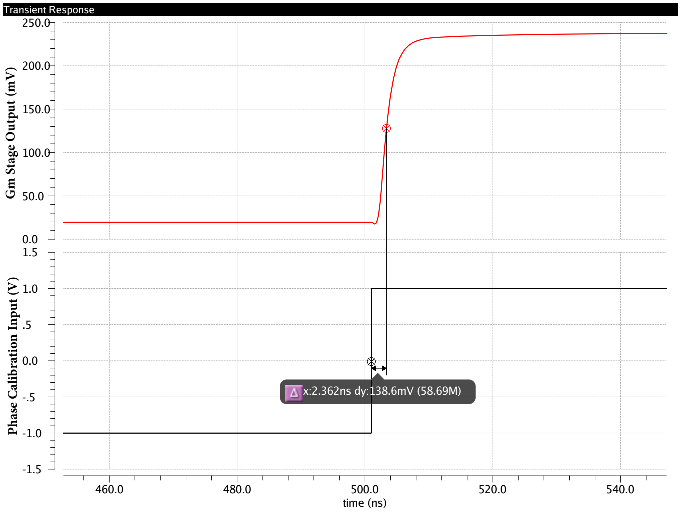


Figure 5.8: Transient simulation result showing the delay from the input of phase calibration IDAC to Gm-stage output

the voltage-to-frequency relationship will be non-linear. The ratio of V_{CCO} to I_{CCO} represents the input impedance of the CCO or Z_{CCO}. In principle, as long as Z_{CCO} is non-zero (a practical challenge), F_{CCO} is a non-linear function of I_{IN}. For an ideal capacitor load and ideal inverters, this impedance is inversely related to inverter transconductance ($1/g_M$). Because g_M itself increases with higher current, the impedance Z_{CCO} is inversely related with CCO frequency.

Periodic small signal (AC) simulations of this structure in a 40nm environment agree with this hypothesis, where N = 3,5 and 7 stage oscillators were simulated. The inverters used in this simulation were twice the minimum size standard library inverters. When Z_{CCO} is plotted versus current, we obtain Fig. 5.12; where N =3,5,7 cases all follow the same curve. This behavior is fundamentally defined by the g_M/I_{CCO} ratio of the MOS devices, and it can be observed that CCO impedance and hence non-linearity improves when g_M improves along with the current.

In order to see how this non-linearity manifests, we can take a look at Fig. 5.13 which shows RMS CCO swing as a function of CCO current. As expected, V_{CCO} has a logarithmic relationship with I_{CCO} since its derivative Z_{CCO} exhibits an inverse relationship with g_M. This means that, according to equation 5.1, F_{CCO} is fundamentally a non-linear function of current. This systematic non-linearity is well known in the literature, with an example for another oscillator given in [5].

Looking at figures 5.11, 5.12 and 5.13, we can see that the linearity of the 3-inverter/stage oscillator is slightly worse, but still comparable to a CCO with 5 or higher number of inverters. The 3-stage CCO exhibits higher impedance (and thus non-linearity) for the same operating frequency, but requires higher current and thus has a worse frequency-to-current gain. It also operates at a lower swing compared to 5 or higher stages, making it more suitable for sub-1V operation.

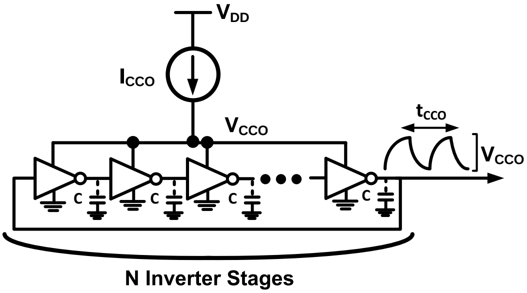


Figure 5.9: Schematic of an N-stage inverter based CCO driven by an ideal current source

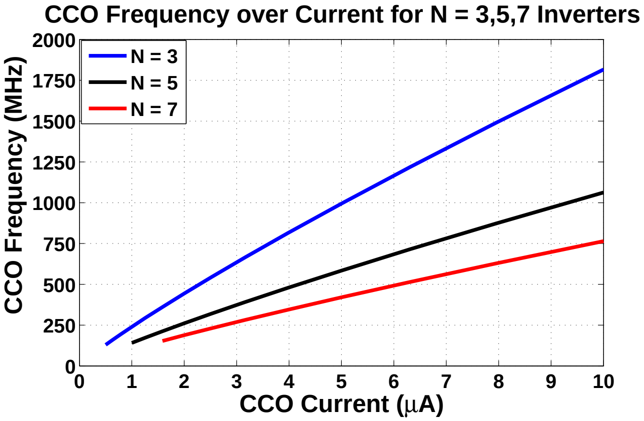


Figure 5.10: CCO Frequency versus current for different number of stages and no load

Therefore, after all these analyses, the 3-stage inverter was chosen for this implementation.

In addition to the systematic non-linearity of the CCO, the MOS transistors themselves add additional non-linearity since their transconductance is a non-linear function of their input. Moreover, parasitic capacitor C, shown in Fig. 5.9, are usually dominated by the gate capacitance of the MOS devices, which are themselves non-linear as well. The capacitive load of the CCO, to be discussed in section 5.4.4 will also influence the CCO behavior. In order to capture all of these effects, single-tone harmonic distortion and two-tone intermodulation tests were done, where the complete VCO (gm and CCO) were driven around the ETF drive frequency. Figure 5.14 shows the power spectral density of the frequency of the CCO when it is driven at 1.125 MHz for the single tone test, and driven at 0.875 and 1.125 MHz for the two-tone test. The peak-to-peak frequency swing on the CCO was limited to 400 MHz for both simulations. The result of the two-tone intermodulation test was used as a sanity check for the harmonic distortion results, and the two results give similar

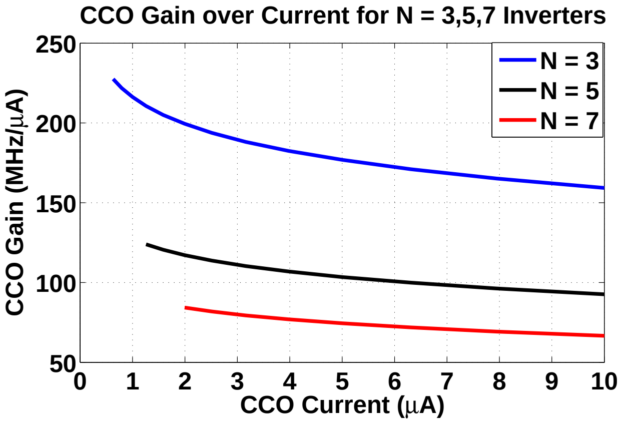


Figure 5.11: CCO Gain over Current for Different Number of Stages and No Load

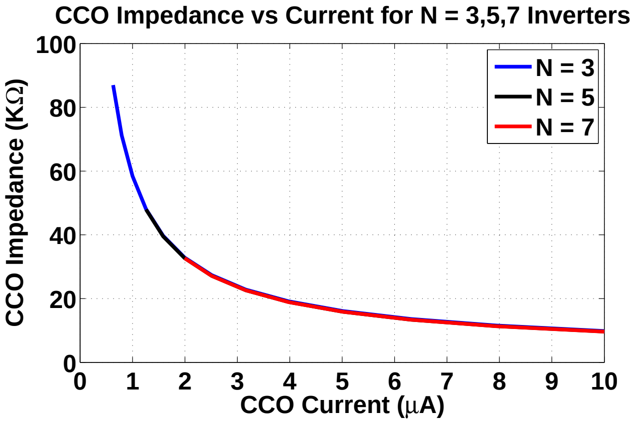


Figure 5.12: CCO Impedance over Current for 3-7 stage oscillators

conclusions about the odd order non-linearity of the VCO.

From Fig. 5.14, we can see that the second harmonic is at -59 dB, while the third harmonic is at -53 dB. Note that this non-linearity is in the amplitude domain. As discussed before in section 3.6.4, a multi-bit PDΣΔM significantly suppresses the non-linearity in amplitude domain. The simulated HD3 of -59 dB results in a negligible phase error of 0.05m^{o} if a 3-bit phase DAC is used.

With the level-shifter load and post-layout extraction, K_{cco} decreases, as shown in Figure 5.15, to 140 MHz/μ A at an operating frequency around 500-600 MHz. Combined with the 1.5 mA/V effective transconductance of the gm-stage, the gain of the gm+CCO combination (VCO) is ~ 200 MHz/mV for the typical corner, close to the desired target specification of 180 MHz/mV.

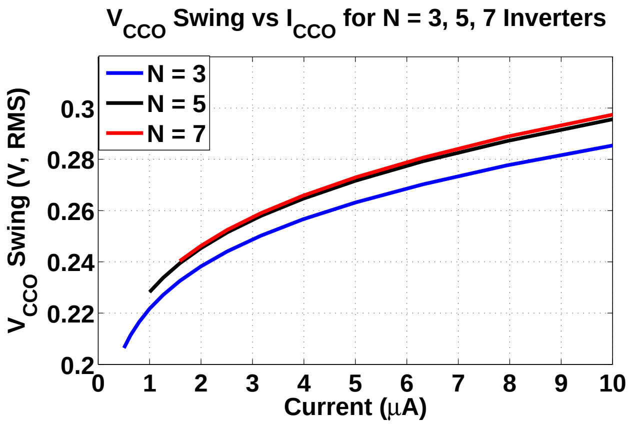
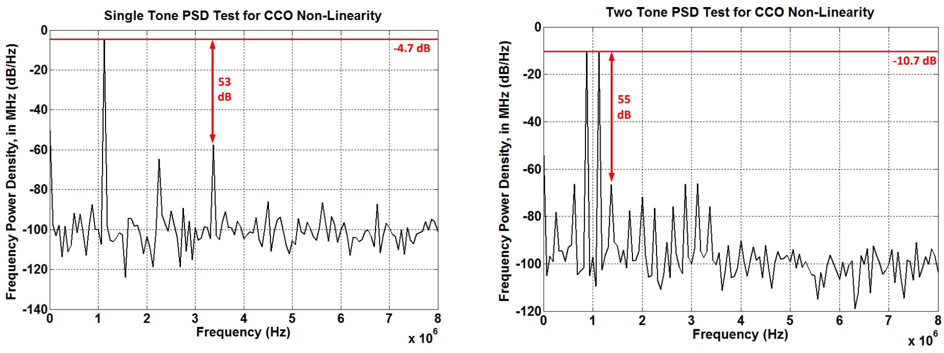


Figure 5.13: CCO RMS Swing over Current for 3-7 stage oscillators



(a) A single tone at 1.125 MHz

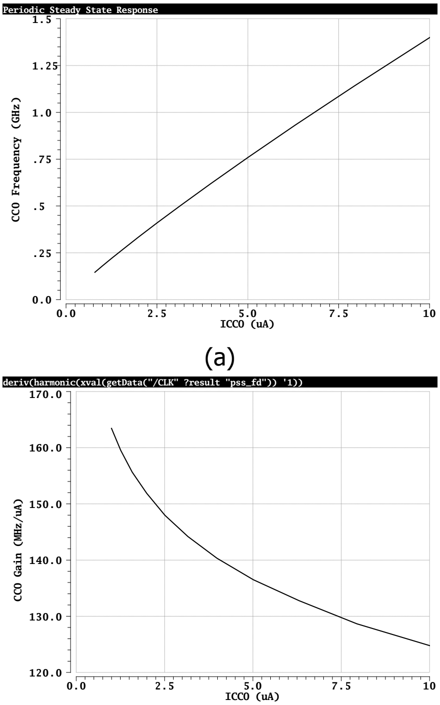
(b) Two tones at 0.875 and 1.125 MHz

Figure 5.14: Power spectral density of ±200 MHz swing on the VCO, for a single tone harmonic distortion and two-tone intermodulation test

5.4.4. Post-VCO Amplifier

The ring VCO discussed in the previous section has an amplitude swing that is not rail-to-rail and is heavily dependent on PVT. In order to reliably interact with the following up/down counter, its output swing must be converted to a rail-to-rail digital signal with a duty cycle $\sim$ 50%. Failure to achieve this could violate the setup or hold time of the counter. Therefore, an amplifier is necessary to convert the VCO output signal into digital domain. Since this amplifier is before the counter, its delay is additive to the ETF signal and hence it must be minimized. Thus, the amplifier should be simple and provide high bandwidth.

An inverter would be ideal for this operation, since it is fast and rail provide a rail-to-rail output. However, its characteristics depend too much on process and



(a)

(b)

Figure 5.15: CCO frequency over current (a) and frequency gain over current (b), after post-layout extraction

temperature. This causes the duty cycle of its output to vary significantly between corners and temperature. A more reliable approach is to use a differential pair coupled with a dummy inverter (of the CCO), as shown in Fig. 5.16. The amplifier provides gain and its output approaches rail-to-rail for all PVT cases. This allows the following inverters to reliably generate a rail-to-rail digital signal. The amplifier consumes 50 μA while exhibiting a bandwidth of 1.8 GHz, which is fast enough to make its contribution to ETF delay negligible.

In order to guarantee reliability, the gm-stage, CCO and the post-CCO amplifier were simulated over supply voltage and temperature. The duty cycle of the CCO frequency over temperature, from 0.9 to 1.2V supply voltage is shown in Fig. 5.17. The duty cycle ranges between 50-62%, which is acceptable for the up/down counter.

5.4.5. Up/Down Counter

The up/down counter acts as the phase demodulator and integrator of the PDΣΔM. Its size is determined from the wrap-around requirements of the modulator as mentioned in section 3.6.2, but its maximum operating frequency and power consumption can still constrain the maximum CCO frequency. It was synthesized via the

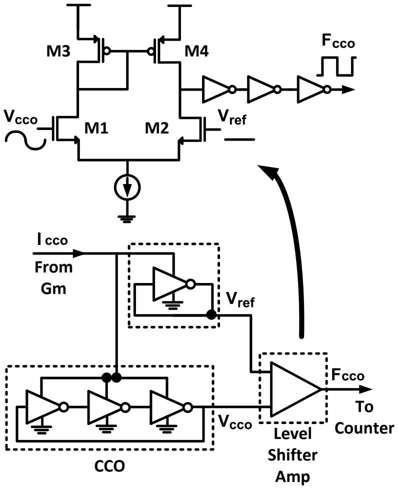


Figure 5.16: Schematic of the post-CCO differential amplifier and how it interfaces with the CCO

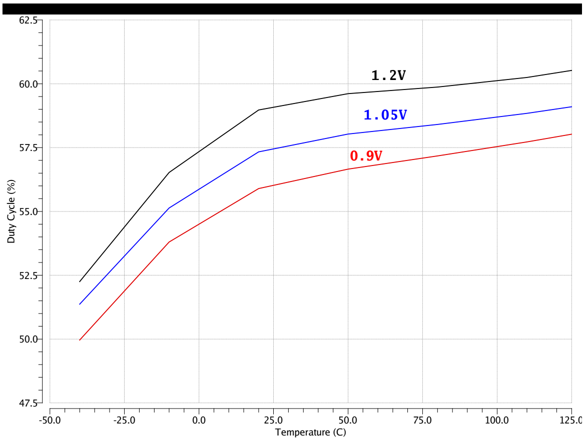


Figure 5.17: CCO frequency's duty cycle over temperature, after the post-CCO amplifier converts it to a digital signal

standard digital flow to simplify the design effort. A custom design for a 0.16μm CMOS process, with a gray code pre-scaler to save power and area, has been demonstrated in [6]. However, the benefit of technology scaling makes this custom design step unnecessary. Operating up to 1 GHz for supply voltages down to 0.9V; the synthesized counter consumes only 0.3 mW power for a nominal frequency of 500 MHz.

The maximum operating speed of the laid out counter was tested over voltage, corners, temperature, and the variable duty cycle of CCO frequency. For stress testing the counter, a constant frequency signal was applied to the counter at the slowest corner, 0.9V supply voltage and -40^o C. The counter was alternated between up and down modes to sweep all possible combinations. For a 0.9V supply, the counter operates for frequencies up to 700 MHz; while at 1V it can operate up to 1 GHz. This is shown in Fig. 5.18, which shows correct operation for 700 MHz at 0.9V.

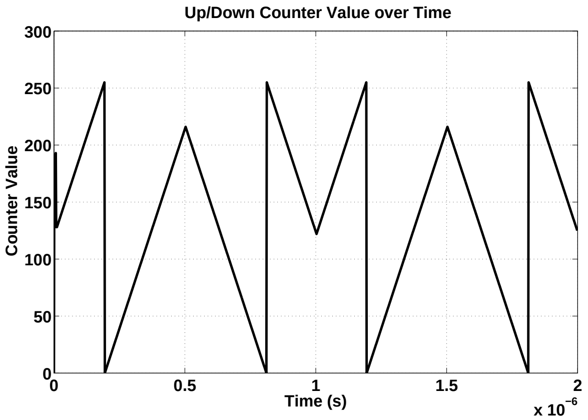


Figure 5.18: Transient simulation results that shows correct operation of the up/down counter for 700 MHz clock, 0.9V supply -40^{o}C temperature and slow corner

5.4.6. Metastability and Sampling

One problem that might arise in the up/down counter is metastability. The CCO frequency and up/down signals present at the counter's input are asynchronous, which means that a metastable condition might occur when these signals have simultaneous transitions. Similar problems have been encountered before in the literature [7], where two asynchronous clock domains must co-exist.

One solution to this problem is to sample the up/down signal by the faster CCO frequency, using an additional flip-flop as shown in Fig. 5.19. As shown in the figure, this small delay causes the up/down signal to be delayed. This changes how the time-domain quantization noise of the counter is calculated and effectively turns the quantization operation of the counter into a round-up operation, as was shown in section 3.6.1.

After the counter, the comparison operation of the digital PDΣΔM is handled by a sampling register. This sampling register has been designed and synthesized as part of the up/down counter.

5.4.7. Digital Heater Drive

The digital heater drive generates the ETF driving and phase reference signals for the PDΣΔM. Timing accuracy of the ETF drive and the phase DAC references can be the bottleneck for sensor inaccuracy, and thus this block is custom designed. This block must also handle the different modes of operation with CCO trim, HDI and low-

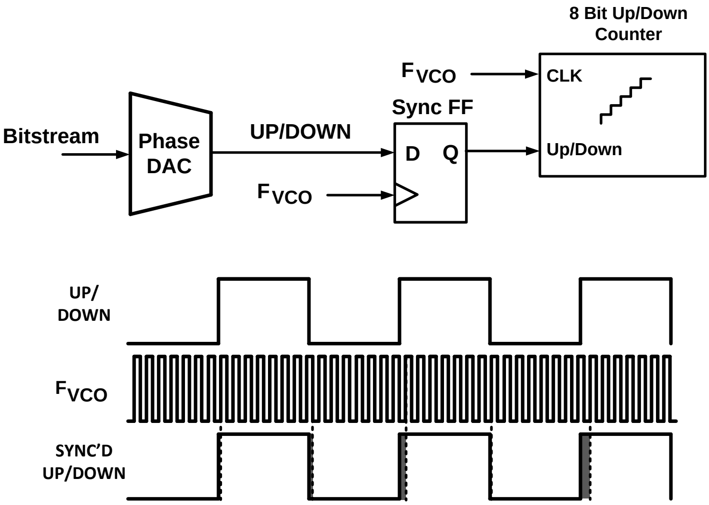


Figure 5.19: Schematic and timing diagram of the up/down re-sampler to avoid metastability

frequency chopping, as well as the phase calibration. Since HDI and low-frequency chopping were not found to be beneficial during measurement, the operation of the heater drive will be explained without them.

Fig. 5.20 shows the schematic of the heater drive. The core of the block is two flip-flops that synchronize the phase references generated by the phase DAC to the 75 MHz synchronization clock F_{SYNC}. Inputs SEL, TRIM and CAL respectively enable sensor operation, CCO trimming mode and calibration modes. Table 5.3 shows the outputs Y, Z, and CAL±that drive the synchronization flip-flops and the calibration IDAC switches in the Gm-stage. For regular operation, the SEL signal must be selected, which directs F_{DRIVE} towards the ETF heater and F_{DAC} to the up/down counter. Both F_{DRIVE} and F_{DAC} are generated by the Phase DAC to be discussed in section 5.4.8.

If TRIM is selected, a high-frequency signal with 50 % duty cycle and frequency $F_{SYNC}/2$ is applied to the ETF to preserve its DC offset (due to self-heating) during CCO trimming. Up/down counting is disabled to use the counter as a fixed frequency divider. If CAL is selected, the CAL$\pm$ signals (driving the phase calibration switch in Fig. 5.5) are driven by F_{DRIVE} and its inverse. This switches the phase calibration IDAC between the two terminals of the ETF, thus generating a square wave signal. Similar to the CCO trim mode, the $F_{SYNC}/2$ signal is applied to the ETF to preserve its offset during calibration.

After the synchronization flip-flop, tapered inverters are used to buffer the ETF drive signal, which drives a 10Ω NMOS switch. Together with the PMOS switch that enables HDI mode (not used in this work), the NMOS switch controls the operation of the ETF heater. Instead of the return-to-float drive of the ETF used in Chapter 4, the proposed heater drive features return-to-Vdd where the ETF heater is reset

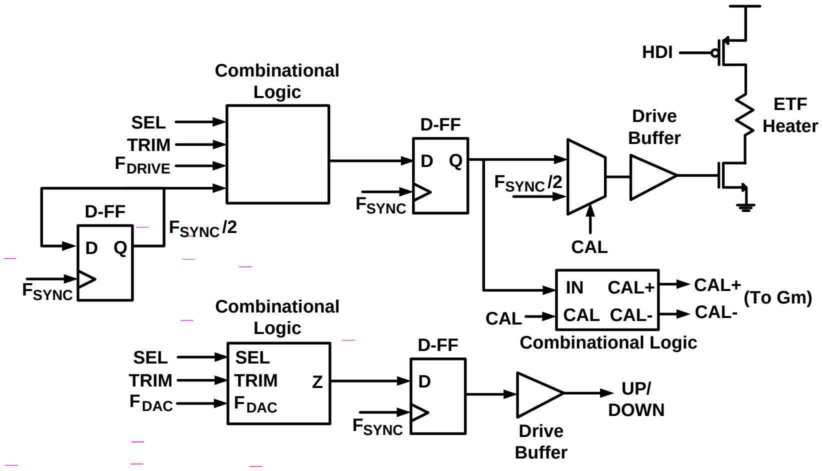


Figure 5.20: Schematic of the digital heater drive

Table 5.3: Combinational Logic Table of the Heater Drive

SEL	TRIM	CAL	Y	Z	CAL+	CAL-
0	X	X	0	0	0	0
1	0	0	F_{DRIVE}	F_{DAC}	0	0
1	0	1	F_{DRIVE}	F_{DAC}	F_{DRIVE}	$\overline{F_{DRIVE}}$
1	1	0	$F_{SYNC}/2$	0	0	0

to the supply rail when the heater drive is turned off. This was done to avoid using large and relatively slower PMOS switches, which can occupy large area and require power-hungry buffers to drive.

Any mismatch in the delays of the synchronization flip-flops, the following logic and the switches themselves will cause a phase error and increase inaccuracy. Phase calibration can detect the error due to the flip-flops, but not the buffers or the switches themselves; which are left as residual error sources. Therefore, the drive buffer is also replicated for the up/down signal path to match the delay of up/down and drive signal paths. The residual error of the heater drive can be estimated from the absolute timing delay between F_{SYNC} and the voltage pulse over the ETF heater, both of which are shown in Fig. 5.21. Assuming 20 % mismatch corresponds to the 3σ spread, the simulated 160 ps delay results in an inaccuracy of 14 m ^{o}s or estimated 0.1 ^{o}C.

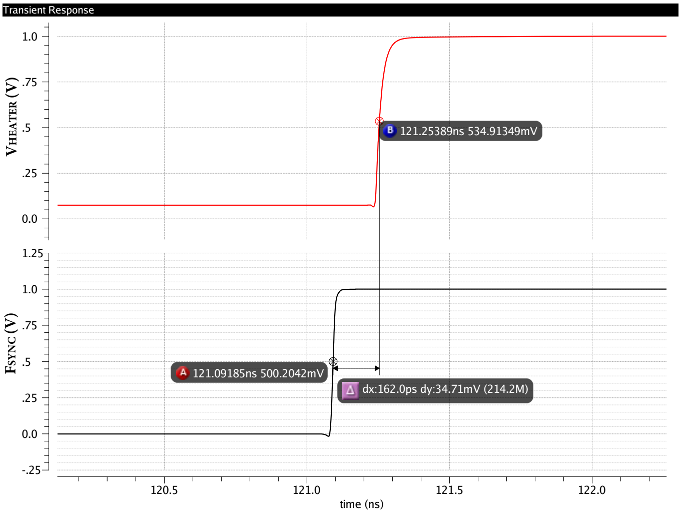


Figure 5.21: Timing delay between the sychronization clock and voltage pulse over the ETF heater

5.4.8. Phase DAC

The Phase DAC generates the ETF drive signal F_{DRIVE} and the up/down signal from the high-frequency reference clock denoted as F_{SYNC}. The DAC provides these by dividing F_{SYNC} by 64 and using an array of flip-flops to shift the result with steps of $1/F_{SYNC}$. The schematic of the Phase DAC is shown in Fig. 5.22. The reference frequency is divided by a ripple counter; which provides the intermediary frequencies. Eight flip-flops act as the delay elements and depending on the modulator output (bitstream), one of these flip-flops are selected to drive the up/down signal. Additional delay elements exist due to the inherent 90° delay related to phase demodulation, as described in section 3.5. To test the feasibility of the DAC with different phase steps, the DAC can be programmed via the signal MODE to operate with a phase LSB corresponding to 5.625 or 11.25°. Since the number of delay elements is fixed, the DAC respectively covers a range of 45° or 90°.

During phase calibration, the ETF drive signal is purposely delayed to act as the reference phase. This circuit is intended to give a phase shift of 22.5° under all conditions.

5.5. Top Level Implementation

With all the proposed circuits, the complete TD sensor is as shown in Figure 5.23. The sensor requires a precision reference clock and some optional programming bits to generate its bitstream output.

The proposed design was laid out and fabricated in TSMC 40nm CMOS tech-

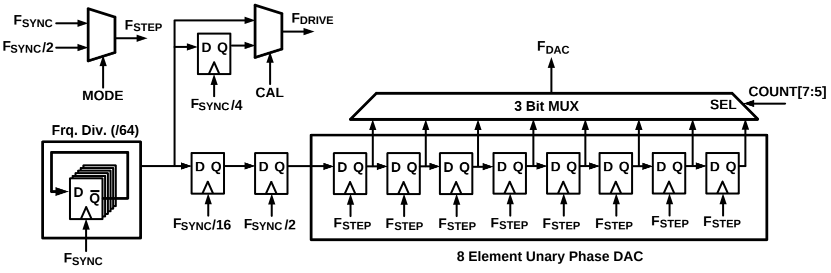


Figure 5.22: Schematic of the Phase DAC generating the ETF drive and up/down signals from an accurate frequency

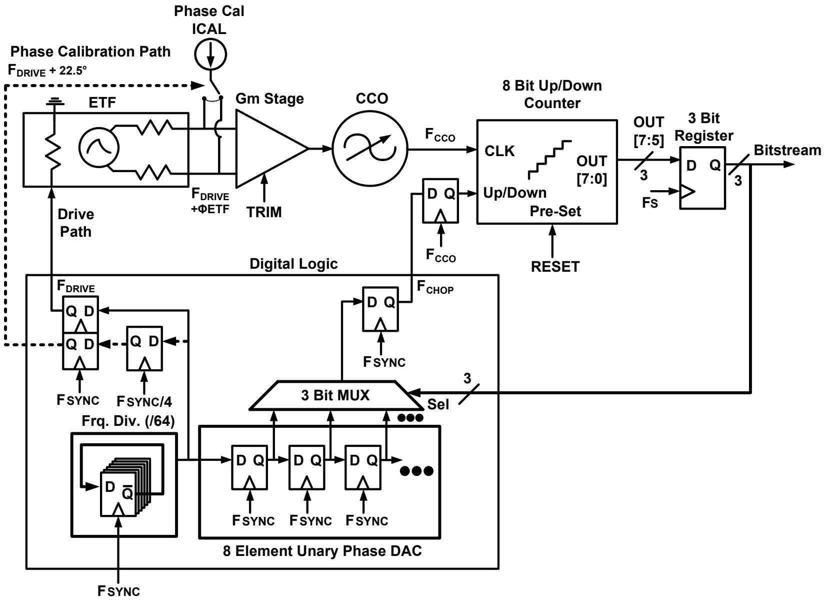


Figure 5.23: Complete block diagram of the TD sensor

nology. The layout of a single TD sensor is shown in Figure 5.24. The drawn dimensions are 30 μm by 68 μm, which results in a silicon area of 1650 μm^2 after a 10% optical shrink.

Since the area of a single sensor is small, an array of 24 sensors was implemented. The majority were used to test various ETF structures. On each chip, the list of implemented ETFs are:

1. 6x ETFs with s = 2 μm and N+ Diffusion Heaters
2. 6x ETFs with s = 3.3 μm and N+ Diffusion Heaters

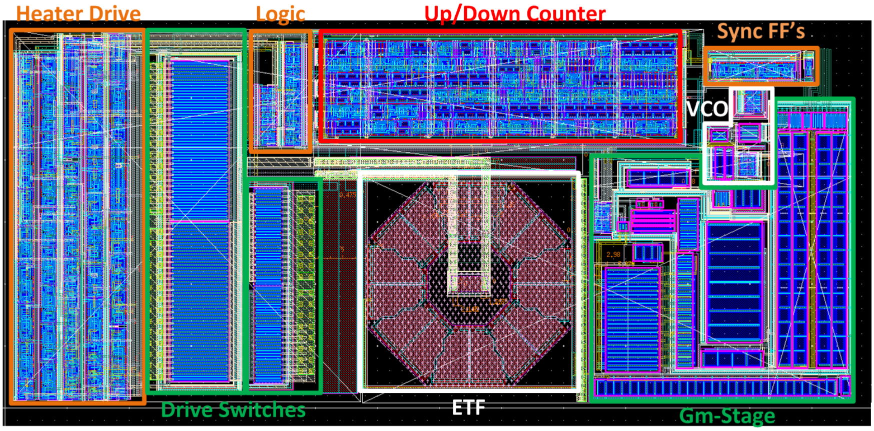


Figure 5.24: Layout of an individual sensor

3. 3x ETFs with s = 2 μm and P+ Diffusion Heaters
4. 3x ETFs with s = 3.3 μm and P+ Diffusion Heaters
5. 3x ETFs with s = 3.3 μm and N-Well Heaters
6. 3x Poly-Poly Oxide ETFs

ETF 6 is not intended for thermal management applications and thus is out of the scope of this work. The resistance of the n-well heater of ETF 5 was significantly higher than expected, and its resolution was too poor; so its measurement results are not reported in this thesis. ETFs 3 and 4 were taped out to test the difference between N+ and P+ diffusion heaters. It was found that ETFs with P+ diffusion heaters have slightly better SNR and energy efficiency, since their heater resistance is lower (114Ω for P+ compared to 189Ω for N+). However, no other difference was found between ETFs with N+ and P+ diffusion heaters, so the measurement results of ETFs #1 and 2 (denoted as s=2 μm and s=3.3 μm respectively) will be mainly discussed in the following section.

5.6. Measurement Results

This section describes the measurement results of the prototype chips containing the TD sensor arrays. The chips were packaged in ceramic DIP28 and plastic SO28 packages, and tested from -40 to 125 °C in a controlled temperature environment. During the measurements, it was observed that using low-frequency chopping (as was done in Chapter 4) does not improve accuracy. This is because an up/down counter behaves like an ideal demodulator and does not introduce residual error like in an analog chopper [8]. It was also found that HDI also does not improve accuracy.

5.6.1. Measurement Setup

Similar to section 4.6.1, the prototype TD sensors were tested on a PCB board with slots for four chips. A Cyclone IV FPGA interfaces with the chips and an off-board NI6537B DAQ card is to transfer the bitstream of sensors to a PC-LABVIEW environment. There, the bitstream is decimated and the results are stored. The same 50ppm 75 MHz MEMS clock source, as in Chapter 4, was used as the timing reference [9]. Figure 5.25 shows the block diagram of the measurement setup.

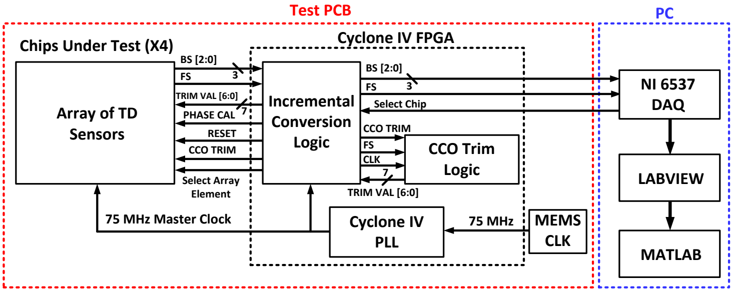


Figure 5.25: Block diagram of the measurement setup used for testing the prototype chips

The timing diagram of the measurement, as determined by the FPGA code, is shown in Fig. 5.26. First, the up/down counting action is disabled, and the CCO is forced into the trim mode, which is described in the next section. The trim mode ends when the CCO reaches the target frequency, and then the reset phase begins. In this phase, the up/down counter is set to its mid value (10000000) to reset the ΣΔ integrator, and then the loop is allowed to run for 1088 cycles. The first 64 cycles are ignored by the FPGA due to settling action of the modulator, as described in section 3.6.3; and when the VALID signal is set high, the subsequent 1024 cycles are forwarded to the PC to be decimated.

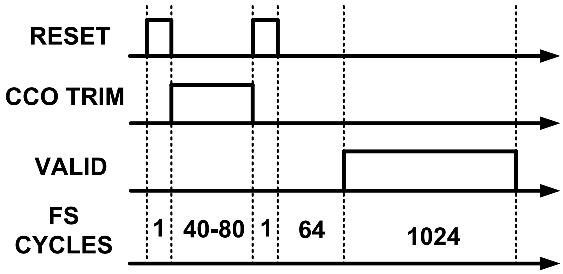


Figure 5.26: Timing diagram of the incremental conversion used for measurement

The complete conversion takes between 1130 to 1170 cycles, depending on nominal CCO frequency. For a drive frequency of 75 MHz / 64 = 1.172 MHz; the conversion time is roughly 1 ms.

5.6.2. CCO Trim Algorithm

The CCO is trimmed via a ramp algorithm that initially gives enough offset to the trim IDAC to stop the CCO. Then, the IDAC setting is slowly increased to turn on the CCO and then increase its frequency until it approaches the target.

During trimming, the CCO's frequency is forwarded to the FPGA via the FS (sampling) signal pin. The FPGA monitors this frequency by counting the number of rising edges received within a reference period of 0.8 μ s. After comparing this count with a target value, the FPGA algorithm then decides whether to increase the IDAC value or not. If the target frequency is reached, the IDAC value is left as is; otherwise, the IDAC value is updated and loaded to the chip. The block diagram of the sensor operation during CCO trimming operation is shown in Fig. 5.27. In effect, this simple trimming algorithm searches for a minimum CCO frequency that guarantees a suitable K_{VCO} for low quantization noise, while avoiding too high frequencies.

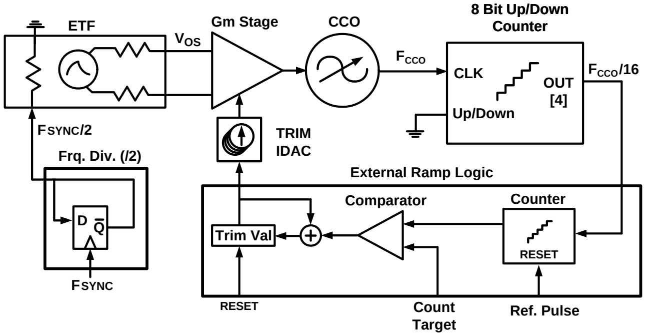


Figure 5.27: Block diagram of the TD sensor during CCO Trim

5.6.3. ETF Phase-to-Temperature Curves

The phase-vs-temperature behavior of the ETFs were characterized by the same method applied in section 4.6.2. The phase shift of ETFs were measured at nine temperature points: -40, -20, 5, 25, 45, 65, 85, 105 and 125 °C, with 100 conversions per temperature and ETF. An aluminum block in thermal contact with the chips and an embedded pt100 inside the block was used as the thermal reference. The master curve was generated from a fifth-order polynomial to approximate the $T^{0.9}$ behavior of the ETFs.

Since readout errors add additional phase shift to the master curves, they must first be monitored and removed via phase calibration. Therefore phase calibration results were also obtained at nine temperature points. From these temperature points, a calibration master curve and a second-order polynomial fit of the phase calibration results were also generated. The behavior of phase calibration results will be shown further on, in section 5.6.4.

The temperature can be roughly estimated via the first, untrimmed ETF master curve, and one-time phase calibration results can then be further corrected over temperature using the previously generated calibration master curve. The corrected phase calibration error can then be subtracted from the ETF phase shift to obtain the calibrated phase vs. temperature curves of ETFs, which is shown in Fig. 5.28. For each phase calibration modes (no phase calibration, one-time or continuous phase calibration), different master curve and fitting polynomials are generated and used.

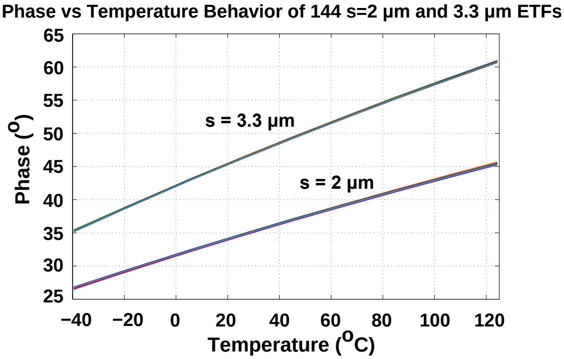


Figure 5.28: Phase-vs-Temperature behavior of 144 ETF1 and 2 sensors, with one-time phase calibration, from -40 to 125 °C

During the measurements, s3.3 μm ETFs required a wide phase DAC range to avoid wrap-around at high temperatures, and were tested with 11.25 ° phase LSB mode. However, s2 ETFs with a higher signal benefited from the reduced integrator swing and relaxed wrap-around requirements of the 5.625 ° LSB mode. Other than the aforementioned wrap-around concerns, no master curve changes or resolution improvements were seen between the two modes.

5.6.4. Phase Calibration

Figure 5.29 shows the phase error recorded during the phase calibration step for 144 sensors (s3.3 ETF) from -40 to 125 °C. The phase error of all 144 readouts is roughly constant over temperature. However, a slight parabolic curvature can be observed, which can be compensated by the second order polynomial of the phase calibration master curve. For this measurement, the phase calibration IDAC strength was set to maximum. Subsequent measurements with different IDAC strengths slightly change the master curve, as shown in Fig. 5.30. During nominal operation, to achieve maximum resolution, the highest calibration IDAC setting was chosen for all ETFs.

Figure 5.31 shows the mean phase calibration error over CCO frequency for s2 and s3.3 ETFs at 25 °C. From the graph, it appears the readout error/delay is a function of $1/F_{CCO}$; as if the CCO is introducing a delay. What we are observing here is the change in the gm-stage's bandwidth as a function of the CCO's IDAC trim value, as shown before in Fig. 5.7. The presence of this error means that the

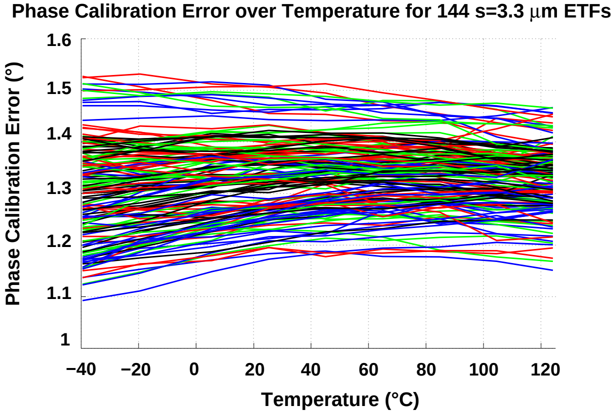


Figure 5.29: Phase error recorded via phase calibration, for 144 sensors employing ETF1, from -40 to 125 °C

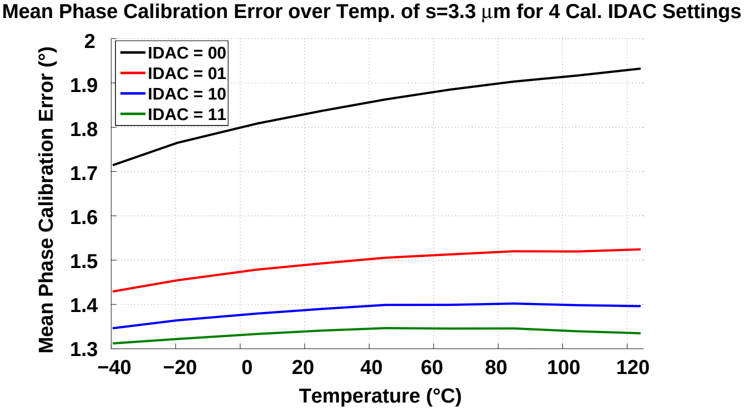


Figure 5.30: Phase calibration master curves for different IDAC strengths, for sensors employing ETF1, from -40 to 125 °C

sensor’s accuracy is affected by spread on the CCO frequency and, the trim IDAC current.

For continuous phase calibration, the mean error values in Fig. 5.29 have been used; while for one-time phase calibration, only the mean values at room temperature have been utilized. The difference between the two methods are too small to be observable for master curves, but they can affect accuracy. Therefore, slightly different master curve coefficients were obtained and stored for continuous and one-time phase calibration. The fifth-order polynomial coefficients that describe ETFs 1 and 2 (after one-time calibration) are shown in Table 5.4.

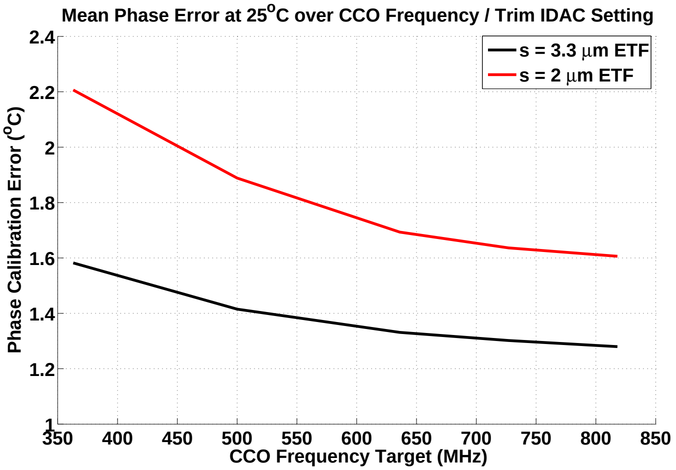
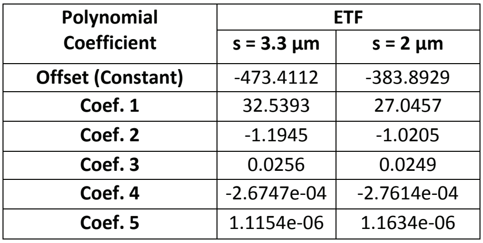


Figure 5.31: Mean phase calibration error at 25 °C, for s=3.3 and s=2 μm, over CCO frequency

Table 5.4: Master Curve fifth order polynomial coefficients that translate phase shift into temperature



Polynomial Coefficient	ETF	
	s = 3.3 μm	s = 2 μm
Offset (Constant)	-473.4112	-383.8929
Coef. 1	32.5393	27.0457
Coef. 2	-1.1945	-1.0205
Coef. 3	0.0256	0.0249
Coef. 4	-2.6747e-04	-2.7614e-04
Coef. 5	1.1154e-06	1.1634e-06

5.6.5. Resolution and Long-Term Stability

After the ETF's master curves are defined, the sensor's noise performance can be obtained. At room temperature and a conversion rate of 1kSa/s, the PSD of the sensor's 3-bit output, shown in Fig. 5.32, exhibits a thermal noise floor corresponding to a resolution of 0.36°C (RMS) for s=3.3μm and 0.24°C (RMS) for s=2μm ETFs. The noise-floor is flat up to tens of kHz, allowing a flexible trade-off between resolution and conversion time. The power consumption of each sensor is 2.5 mW, where 88% of the power (2.1 mW) is dissipated inside the ETFs.

During resolution measurements, the average resistance seen by the ETF heater supply was found to be 245Ω compared to the 188Ω expected from the layout. The difference of 57Ω results from on-chip, package and PCB parasitics, as well as

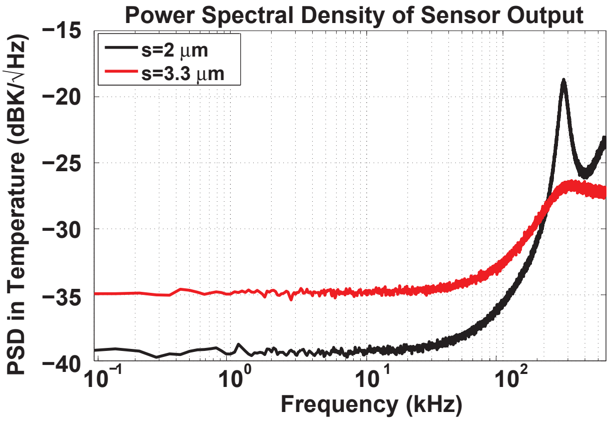


Figure 5.32: PSD of the sensor's bitstream in temperature (8 million points, Fs = 1.17MHz)

heater drive switches. During post-layout extractions, only 35Ω on-chip resistance was estimated, and it can be assumed that the remaining 22Ω s is due to bond wires, packaging, and PCB traces.

Due to this additional resistance, the ETF heater works at only 76% power efficiency. This degradation, as well as the 16% power reduction (2.5 to 2.1mW) compared to [2], results in a 30% reduction in SNR. These results, considering the 24% efficiency loss, agree well with the predictions made in section 5.4.

Long-term stability of temperature sensors can be important for some applications, and it can also be used to differentiate flicker noise, drift, or measurement setup related error sources. This was a concern since flicker (or 1/f) noise in a small ring-VCO can be significant, and it is important to demonstrate if this can limit the long-term stability of the sensor.

For this purpose, the resolution of the sensor's bitstream was checked over variable conversion time. The result of this measurement with 8 million samples (limited by DAQ memory) is shown in Figure 5.33. The red line shows how the sensor's resolution in phase ($\sigma(\tau)$) improves as the conversion time (τ) is increased. Between 10μs and 1s, the sensor's output follows the dashed fitted line with a slope of -0.5 in the log-log domain, which corresponds to a $\sqrt{\tau}$ relationship. This line corresponds to a thermal noise fit. From the figure, it is visible that the sensor exhibits only white noise down to 1 second conversion time. The 3 σ error corresponding to 100 measurements at 1 kSa/s due to noise is roughly 0.1 °C, which is the ultimate inaccuracy of this measurement setup.

5.6.6. Inaccuracy

In order to evaluate the inaccuracy of the TD sensors, 24 ceramic DIL packaged chips have been characterized over a temperature range of -40 to 125 °C. When operated from a 1.05V supply, the sensors with s = 3.3 μm achieve an untrimmed inaccuracy of ±1.8°C (3σ) as shown in Fig 5.34. This improves to ±1.4°C (3σ) after one-time phase calibration, and to ±0.75°C (3σ) after temperature trimming

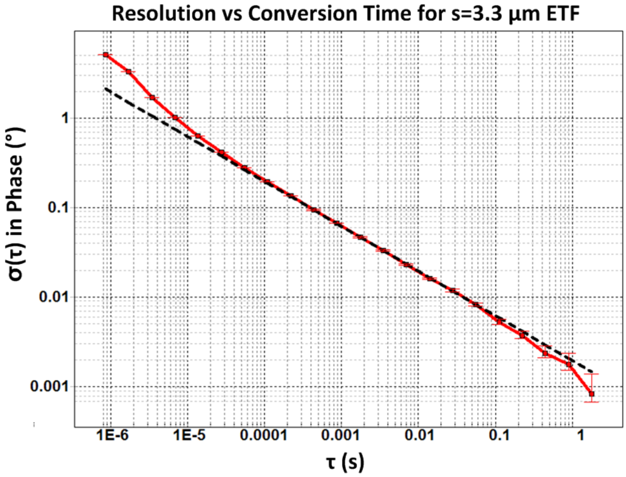


Figure 5.33: Log-log plot showing readout's resolution vs conversion time

at 25°C. Continuous phase calibration improves the untrimmed inaccuracy only marginally but improves trimmed inaccuracy to ±0.5°C (3σ).

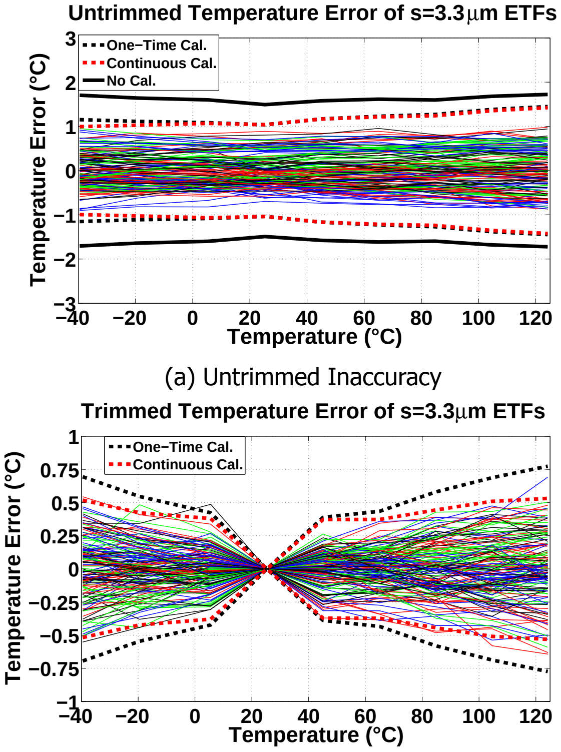
Fig 5.35 shows the untrimmed and trimmed inaccuracy of the smaller s2 ETF. As expected, its improved resolution comes at the expense of accuracy: their untrimmed inaccuracy is ±2.3°C (3σ, 144 samples) after a one-time or continuous phase calibration. After a single-point trim, those values reduce to ±1.05°C (3σ) and ±0.85°C (3σ), respectively.

The sensors have also been verified over 0.9-1.2V analog and digital supply range. At a 0.9V supply voltage, the digital logic and Gm-stage slows down, resulting in worse performance. For s3.3 ETF, the untrimmed inaccuracy becomes ±2.3°C (3σ), and ±1.2°C (3σ) after trimming, while for the s2 ETF, the untrimmed inaccuracy is ±2.5°C (3σ), and ±1.4°C (3σ) after trimming.

The primary inaccuracy source (apart from the ETF) is believed to be from the residual error of nominal CCO frequency, as was mentioned in Fig. 5.31. The boundaries of this error can be estimated by assuming that the mean error on CCO frequency is bounded within one LSB ($\sim$ 100 MHz), corresponding to one LSB of the trim IDAC (0.5 μ A). From figure 5.31, the equivalent phase error is 46 m° for s=3.3 μm and 98 m° for s=2 μm for a nominal CCO frequency of 630MHz and a trim error of 100 MHz. The inaccuracy in temperature is ±0.3 °C for s=3.3 μm and ±0.7 °C for s=2 μm.

5.6.7. Linearity and Ramp Measurements

A ramp temperature test was done to verify that the non-linearity errors present in two-step PDΣΔMs are no longer present when multi-bit feedback is used. For this measurement, the setup was brought to -40 °C first and then the temperature was ramped up slowly to 105 °C, and conversions were made for 24 s=3.3 μm sensors (4 chips) at steps of 50m °C. In order to average out noise, groups of



(a) Untrimmed Inaccuracy

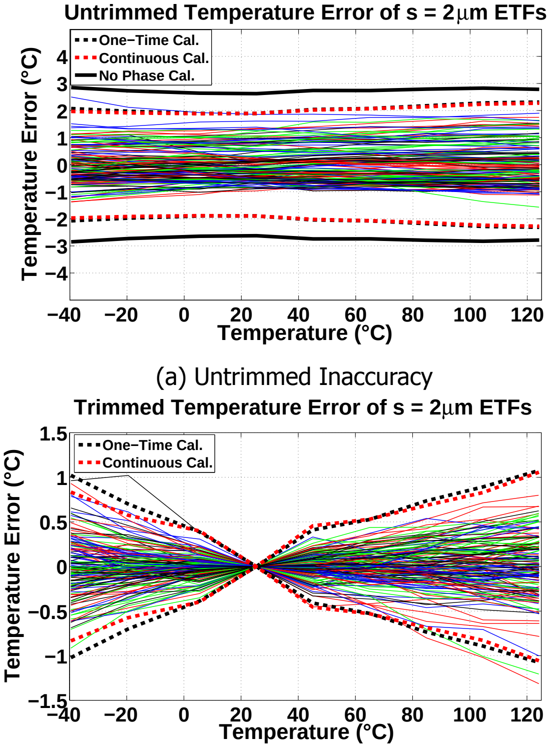
(b) Room temperature Trimmed Inaccuracy

Figure 5.34: Untrimmed and trimmed inaccuracy of 144 s=3.3μm ETFs. Individual lines represent the inaccuracy of each sensor with one-time phase cal., while the bold lines indicate the 3σ limits for no phase cal., one-time phase cal. at 25 °C, and continuous phase cal.

10 conversions were each binned and averaged into data points corresponding to 0.5 °C with a standard deviation (σ, RMS) of 0.11 °C for each point. The master curve of this ramp measurement was checked against the polynomial described in Table 5.4, and it was seen that the master curves align with an error of ±1°C over temperature, which is theorized to be due to a temperature gradient between the pt100 and ETFs during the ramp measurement.

Fig. 5.36 shows the temperature error for 24 sensors in one ramp measurement, when compared to the master or mean curve. It is interesting to note that the visible 0.11 °C (RMS) noise does not significantly degrade the accuracy, and the measured inaccuracy is comparable to results obtained from Fig. 5.34a.

The mean non-linearity or INL of the sensor can be distinguished by comparing the temperature of the ETFs to the 50m°C/point ramp of the oven. Fig. 5.36 shows this difference, and a slow settling behavior can be seen up to 10 °C. This settling is due to the formation of a thermal gradient inside the oven when the temperature is slowly ramped up. Unlike the two-step conversion method, no non-linear behavior



(a) Untrimmed Inaccuracy

(b) Room temperature Trimmed Inaccuracy

Figure 5.35: Untrimmed and trimmed inaccuracy of 144 s=2μm ETFs. Individual lines represent the inaccuracy of each sensor with one-time phase cal., while the bold lines indicate the 3σ limits for no phase cal., one-time phase cal. at 25 °C, and continuous phase cal.

or glitches can be observed and the linearity error from -40 to 125 °C is confirmed to be less than 0.1 °C.

5.6.8. Mechanical Stress and Plastic Packages

In order to test the effects of mechanical stress on the ETFs, 15 chips were packaged in SO28 plastic packages and tested over temperature. The master curve of ETFs in both ceramic and plastic packaged chips were tested and compared. Due to the change in thermal properties of the package (plastic versus ceramic), a temperature shift was observed on the reference pt100 sensor when compared to previous results. This was verified by comparing the mean temperature difference of the pt100 versus the oven temperature with S028 and ceramic packages. Even when repeated over multiple runs, it was found that the pt100 temperature tracks oven temperature well when the same (plastic or ceramic) package type is tested. The variance of pt100 temperatures over different oven runs was found to be less than 0.1 ^oC. However, when the package type changes from ceramic to

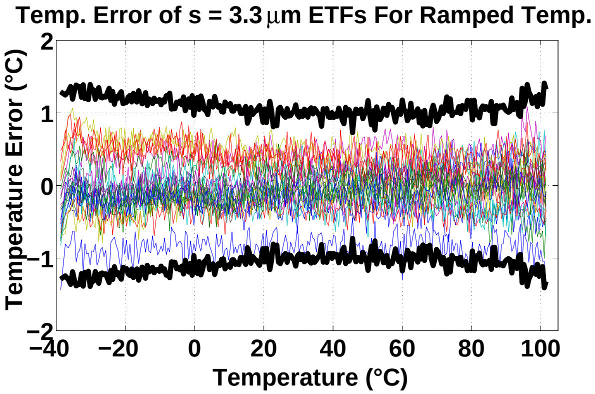


Figure 5.36: Temperature error of 24 s=3.3 μm ETFs to a ramped temperature test. Bold lines indicate 3σ limits.

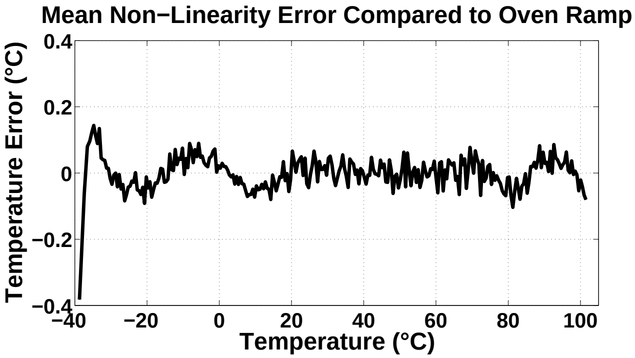
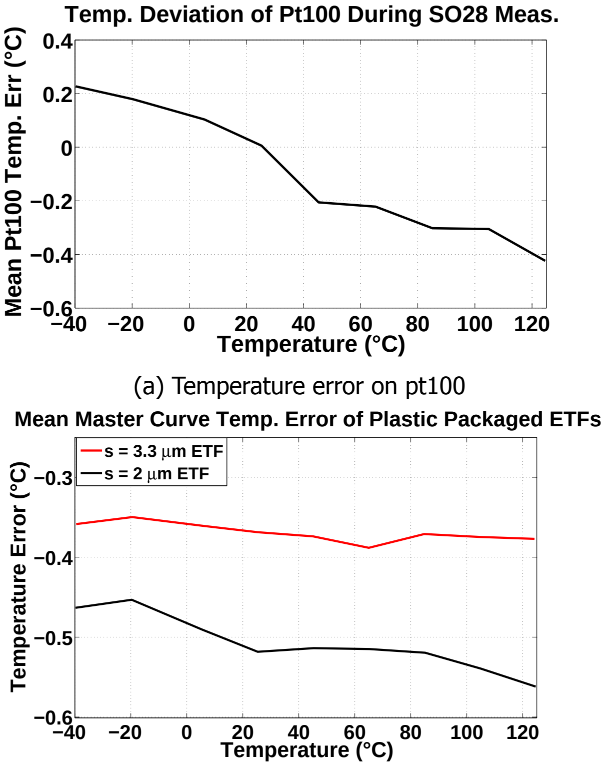


Figure 5.37: Non-linearity error between oven ramp and mean sensor output over temperature

SO28; a shift in pt100 temperature data was observed for the same target oven temperatures.

Simultaneously, a constant temperature shift was observed on the master curves of plastic and ceramic packaged ETFs. The mean of both of these errors over temperature is shown in Fig. 5.38. When these two error curves are added, a clear PTAT error curve emerges. Self-heating of the SO28 packages is not expected to be the cause, since SO28 packaged sensors under-estimate the temperature with respect to ceramic packaged ones, even though plastic packages exhibit higher thermal impedance and are expected to run slightly hotter. However, this under-estimation of temperature is in line with an increase in thermal-diffusivity under compressive strain [10]. Therefore, it is hypothesized that this PTAT error is due to mechanical stress on the SO28 packages, as mentioned in section 2.6.3.

SO28 packages also exhibit higher spread, for both ETFs, especially without temperature trimming. The untrimmed inaccuracy degrades to ±2.3°C (3σ) for s=3.3 μm, and to ±3.8°C (3σ) for s2 ETF. One-point temperature trimming im-



(a) Temperature error on pt100

(b) Plastic package mean temperature error

Figure 5.38: Mean temperature variation of the reference pt100 sensor during measurements of SO28 packages, and mean error of the master curve of plastic packaged ETFs compared to ceramic packaged ones.

proves the inaccuracy to ±0.9°C (3σ) for s3.3 ETF and ±1.4°C (3σ) for s2 ETF. The degradation in untrimmed accuracy is much worse, so it can be concluded that a single point trim helps to alleviate the spread due to the variation of mechanical stress. Instead of single-point offset trimming, PTAT trimming (widely used in BJT-based sensors) can be adopted [8]. For this trim, the temperature error measured at 25 °C is assumed to be a function of absolute temperature and scaled to such effect. As shown in Fig. 5.39, the trimmed inaccuracy of s=3.3 μm improves down to ±0.75°C (3σ) after a PTAT trim, which is the original spread in ceramic packages after an offset trim.

Therefore, it can be inferred that PTAT trimming is a better alternative for plastic packaged sensors and that the increased spread due to mechanical stress is indeed a PTAT effect. As a conclusion, despite the sensor's susceptibility to mechanical stress, plastic packaged sensors perform as well as ceramic packaged ones after one-point PTAT trimming.

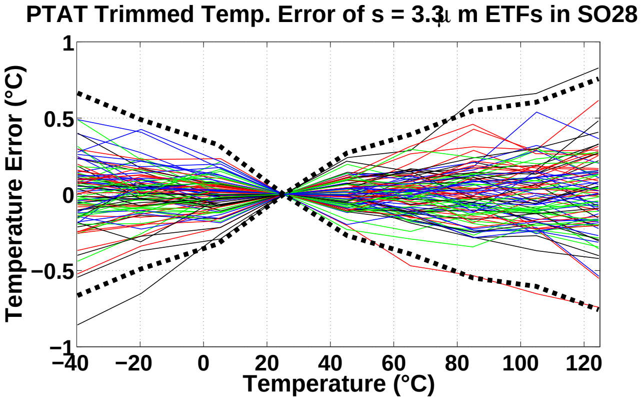


Figure 5.39: PTAT Trimmed Inaccuracy of 90 SO28 packages s = 3.3 μm ETFs

5.7. Conclusion

Table 5.5 summarizes the performance of both sensors (with s=3.3μm and s=2μm ETFs) and compares them to other compact state-of-the-art sensors intended for thermal monitoring applications or sensors implemented in advanced processes. The proposed sensor (with the s=3.3μm ETF) is the most accurate and the smallest, except for a sensor that requires a precise external voltage reference (which is not included in the size reported) [11] .

It also has the second lowest operating supply voltage (0.9V), which is mainly limited by the up/down counter. Compared to TD sensors implemented in more mature technologies [1] [6], it achieves roughly 1.5x better resolution and 2x more accuracy, while requiring about 2x less area. These results demonstrate that TD sensors scale well in nanometer CMOS, and can be used to realize accurate, low-voltage, and compact temperature sensors for thermal monitoring.

References

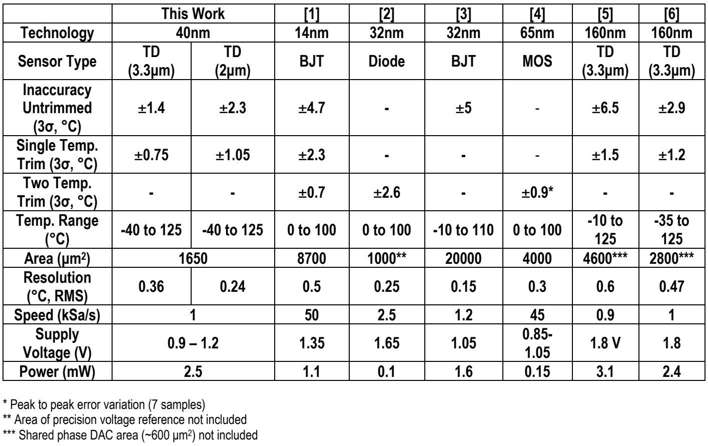
[1] R. Quan, U. Sonmez, F. Sebastiano, and K. Makinwa, "A 4600 μ m2 1.5 °c (3 σ) 0.9ks/s thermal-diffusivity temperature sensor with VCO-based readout," in *IEEE ISSCC Dig. Tech. Papers*, Feb 2015, pp. 488–450.

[2] U. Sonmez, R. Quan, F. Sebastiano, and K. Makinwa, "A 0.008 mm2 area-optimized thermal-diffusivity-based temperature sensor in 160nm cmos for SoC thermal monitoring," in *ESSCIRC*, Sept 2014, pp. 395–398.

[3] J. Shor, K. Luria, and D. Zilberman, "Ratiometric bjt-based thermal sensor in 32nm and 22nm technologies," in *IEEE ISSCC Dig. Tech. Papers*, Feb 2012, pp. 210–212.

[4] M. H. Perrott. CppSim System Simulator Package. [Online]. Available: http://www.cppsim.com

Table 5.5: Performance summary and comparison with previous works



	This Work		[1]	[2]	[3]	[4]	[5]	[6]
Technology	40nm		14nm	32nm	32nm	65nm	160nm	160nm
Sensor Type	TD (3.3µm)	TD (2µm)	BJT	Diode	BJT	MOS	TD (3.3µm)	TD (3.3µm)
Inaccuracy Untrimmed (3σ, °C)	±1.4	±2.3	±4.7	-	±5	-	±6.5	±2.9
Single Temp. Trim (3σ, °C)	±0.75	±1.05	±2.3	-	-	-	±1.5	±1.2
Two Temp. Trim (3σ, °C)	-	-	±0.7	±2.6	-	±0.9*	-	-
Temp. Range (°C)	-40 to 125	-40 to 125	0 to 100	0 to 100	-10 to 110	0 to 100	-10 to 125	-35 to 125
Area (µm²)	1650		8700	1000**	20000	4000	4600***	2800***
Resolution (°C, RMS)	0.36	0.24	0.5	0.25	0.15	0.3	0.6	0.47
Speed (kSa/s)	1		50	2.5	1.2	45	0.9	1
Supply Voltage (V)	0.9 – 1.2		1.35	1.65	1.05	0.85-1.05	1.8 V	1.8
Power (mW)	2.5		1.1	0.1	1.6	0.15	3.1	2.4

* Peak to peak error variation (7 samples)
** Area of precision voltage reference not included
*** Shared phase DAC area (~600 µm²) not included

[5] G. Taylor and I. Galton, "A Mostly-Digital Variable-Rate Continuous-Time Delta-Sigma Modulator ADC," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 12, pp. 2634–2646, Dec 2010.

[6] J. Angevare, L. Pedalà, U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "A 2800-um2 Thermal-Diffusivity Temperature Sensor with VCO-based Readout in 160-nm CMOS," in *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov 2015, pp. 1–4.

[7] R. B. Staszewski, C.-M. Hung, K. Maggio, J. Wallberg, D. Leipold, and P. T. Balsara, "All-digital phase-domain tx frequency synthesizer for bluetooth radios in 0.13 μm cmos," in *IEEE ISSCC Dig. Tech. Papers*, Feb 2004, pp. 272–527 Vol.1.

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

[9] *Low Jitter Pin Configuration CMOS Output 3.2 x 2.5 x 0.85 mm Ultra Miniature Pure Silicon Clock Oscillator*. [Online]. Available: http://www.abracon.com/Oscillators/ASEMCC.pdf

[10] X. Li, K. Maute, M. L. Dunn, and R. Yang, "Strain effects on the thermal conductivity of nanostructures," *Phys. Rev. B*, vol. 81, Jun 2010.

[11] G. Chowdhury and A. Hassibi, “An On-Chip Temperature Sensor With a Self-Discharging Diode in 32-nm SOI CMOS,” *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 59, no. 9, pp. 568–572, Sept 2012.

6 Conclusion

This chapter summarizes the impact of the work described in Chapter 1-5. It also provides a general overview of the main findings of this thesis, its original contributions, other applications where this work might be applied, and gives guidance to future work on ETFs or TD-based temperature sensors.

6.1. Main Findings

The following gives a summary of all the major findings/discoveries that are discussed in this work, specifically for TD-based temperature sensors. The original publications that constitute parts of this thesis are cited separately and relevant chapters are also mentioned.

- For the first time, quasi-ballistic thermal transport is successfully modeled for ETFs, and verified with silicon results. Good agreement was found between the quasi-ballistic model developed in Chapter 2, measurements of compacts ETFs in Chapters 4 and 5, and recent experimental work on short-distance (<10μm) heat transport in bulk silicon [1][2][3]. These results improve our understanding of heat transport in silicon and pave the way for the design of even smaller ETFs that can achieve better resolution.

- ETF scaling from 160nm down to 40nm CMOS has been investigated [4]. It was found that ETF accuracy does not scale proportionally with process feature size (see section 2.6.1). The untrimmed inaccuracy of the same ETF improves from ±2.3 ^{o}C in160nm CMOS, to only ±1.4 ^{o}C in 40nm. Despite this limitation, the impact of scaling is clearly non-negligible and better performance can be expected in more advanced process nodes.

- The VCO-based PDΣΔM introduces additional quantization noise in the time-domain (see section 3.6.1), due to its edge-sampling property [7]. It is shown that this problem can be alleviated by increasing the gain of the VCO, and by properly sizing the up-down counter.

- It is hypothesized that packaging stress on small ETFs degrades inaccuracy, as discussed in sections 2.6.3 and 5.6.8. However, it is found that after a single-temperature PTAT trim (see section 5.6.8), the absolute temperature inaccuracy is the same for both plastic and ceramic-packaged devices. This is critical for commercial applications, since plastic packages are generally preferred due to size and cost reasons. Single-temperature PTAT trim is a relatively low-cost and commonly practiced solution in the industry, so in terms of packaging there is little concern for mass production of TD sensors.

The achieved performance shows that TD-based temperature sensors are suitable for commercial adoption, especially for thermal management applications. Since they scale, it may be expected that future TD sensors will achieve even better performance. Extrapolating from the 1.6x accuracy improvement between 160nm and 40nm CMOS processes, TD sensors in 20nm planar CMOS or 16nm FinFET process nodes can be expected to achieve inaccuracies less than $\pm 1\ ^{o}$C (3σ, no temperature trim).

The proposed sensors are also compact enough to enable other interesting applications: for example that of trimming other, high-resolution but poor accuracy temperature sensors. MOS or resistor based temperature sensors can exhibit great relative accuracy and temperature resolution but poor absolute accuracy. A compact TD sensor can be placed next to such sensors on the same die, and can be used to trim these sensors without waiting for the chip to thermally settle and hence avoiding expensive trim costs. Relaxing the accuracy requirements on such sensors means that they can be designed for maximum resolution rather than accuracy. This can be done, for example, by using continuous-time rather than incremental ΣΔ modulators, or eliminating chopping/DEM schemes which increases design complexity and current consumption.

Another exciting possibility for ETFs is in heat transport research and experimentation.The quasi-ballistic thermal transport properties of small ETFs can be used to investigate and further understand the thermal properties of silicon and other materials used in IC production. Compared to indirect measurement methods such as the use of laser gratings [1], ETFs can measure the thermal properties of bulk silicon directly, at a significantly lower cost. This can allow researchers to obtain more data, faster.

To summarize, this work demonstrates that there is still great potential for TD sensors in several critical applications: thermal management, pairing up with poor accuracy but high resolution sensors, and heat transport experimentation. Section 6.3 also discusses potential improvements to jitter performance of TD-based frequency references due to the advances made in this work regarding TD sensor resolution, and other possible applications for VCO-based PDΣΔMs.

6.2. Original Contributions

The major original contributions of this work are listed below. The original publications that constitute parts of this thesis are cited separately and relevant chapters are also mentioned.

- A new ETF geometry, called as the Octagonal ETF, is proposed. See section 2.4.3 and results under 4.6.
- Improved, frequency-domain modeling of phase-contour and Octagonal ETFs is presented in section 2.5. This model can account for ballistic transport effects and can be used to calculate the temperature resolution, SNR and inaccuracy of such ETFs.
- Two small ETF designs with critical dimensions of 2 and 3.3 μm are designed as part of this work (sections 2.7.1 and 2.7.2). These ETFs are the smallest implementations in the current literature.
- The impact of lithography, mechanical stress and self-heating on ETF inaccuracy is analyzed in section 2.6. The impact of self-heating and mechanical stress is analyzed for the first time.
- A detailed analysis of VCO-based PDΣΔMs is presented in section 3.6 and [7]. Such PDΣΔMs can be used in other applications [8][9]. This analysis on quantization noise, counter sizing requirements, and non-linearity is intended to guide future designs.
- A simple foreground calibration technique is proposed to improve the absolute accuracy of phase-domain ETF readouts [4]. This technique, called phase calibration, is demonstrated in section 5.3. It is shown in section 5.6.6 that using a single-shot phase calibration improves untrimmed inaccuracy from ± 1.8 ^{o}C to ± 1.4 ^{o}C (both 3σ). Continuously applying phase calibration can improve the room-temperature-trimmed accuracy of the sensor down to ± 0.5 ^{o}C (3σ).
- The first TD sensor design in 40nm CMOS is presented [4] (see Chapter 5). This is the TD sensor that can meet the tough area, speed, resolution and inaccuracy specifications required by thermal management applications.

6.3. Other Applications of This Work

While this work describes TD sensors, similar principles can be used to design TD-based frequency references [10]. Such sensors also use ETFs to build frequency references, and thus the work in Chapter 2 on ETFs is relevant in such applications. In TD-based frequency references, a low-noise VCO is locked to the ETF's phase shift. In order to suppress the VCO's jitter and frequency drift over time, the loop must have significant bandwidth, which also increases the thermal noise contribution from the ETF. The high-speed (1kSa/s) phase-domain readout and the high resolution octagonal ETF geometry, both discussed in this work work are significant advancements compared to previous efforts [10]. For a TD-based frequency reference, the improvement in speed and resolution allows the loop bandwidth to be 100s of Hz rather than a few Hz, and hence suppresses VCO flicker noise significantly. Potentially, this can improve the jitter performance of future TD-based frequency references compared to [10].

PDΣΔM or similar circuit architectures have also been used in a wide variety of applications. Some examples include single-photon avalanche diodes (SPADs) [8], wireless receivers [9], resistor-based temperature sensors [11] and CO_2 sensors [**?**]. The analysis of and design improvements made to PDΣΔMs done in this work can be applied to some of these circuits.

The sensors used in [11] and [**?**] both operate at low frequencies and use analog front-ends, and are hence suitable for miniaturization via the adoption of VCO-based PDΣΔM. Here, resistor or RC-based PDΣΔMs have been used in these works to measure either temperature or the local CO2 level and the readout area is typically comparable to the large area required for precision resistors. The challenge for both circuits is to improve the modulator's quantization noise floor since the required noise floor is significantly (at least 40dB) smaller than an ETF. As described in [7], this can be made possible by adopting a second or higher-order VCO-based PDΣΔM, which significantly reduces the quantization noise at the cost of adding some digital logic. While such higher-order and higher-resolution variants of the VCO-based PDΣΔMs are outside the scope of this work, the analysis of PDΣΔMs in Chapter 3 has been used as stepping stone to model higher-resolution VCO-based PDΣΔMs for SPADs, resistor readouts or wireless receivers [7].

6.4. Future Work

This work could be further improved by doing the following:

- Stress sensitivity of ETFs can be further understood by implementing an on-chip stress sensor and measuring exactly how an ETF responds to mechanical stress. This aspect of TD sensors is not yet well understood.

- Even smaller ETFs can be implemented to improve the energy efficiency of future designs. For better lithographic accuracy, ETFs that rely on the critical masks of nanometer CMOS processes should be used. These masks are typically used for polysilicon, gate and metal-1 layers. In such processes, the requirement to build ETF thermopiles out of inaccurate salicide-protection layers seems to limit the ultimate accuracy of ETFs.

- The proposed sensor can be implemented along with a precision temperature sensor, or another silicon-oxide based ETF to build a temperature-insensitive frequency reference as in [10]. Thanks to the techniques presented in this thesis, such a frequency reference can be extremely compact and can be easily implemented as a combined on-chip temperature sensor and frequency generator.

- As explained before in section 6.3, an accurate, but noisy, TD sensor can be combined with a high-resolution, but inaccurate, temperature sensor. An accurate, compact and digital TD sensor can be used as an absolute temperature sensor to trim its high-resolution counterpart. This trimming can be one-time, continuous, or be done periodically for short amounts of time to limit extra current consumption. The combined temperature sensor would have the high-resolution of the relative sensor and the superior absolute accuracy of the TD sensor.

- Energy efficiency and conversion speed of the readout can be further improved by adopting the second-order VCO-based PDΣΔM presented in [7]. For this implementation, a smaller and more thermally efficient ETF should be used to suppress the ETF's thermal noise contribution.
- The accuracy of a VCO-based readout can be improved by better trimming the CCO frequency over temperature, since this is the dominant source of inaccuracy (as explained in section 5.6.6). Locking the CCO frequency to an external source during the CCO trim procedure, and using a finer trim resolution can be used to avoid this problem.

Last of all, the proposed ETF designs and circuit implementation can be implemented in a FinFET technology to investigate the challenges and opportunities present in a FinFET technology compared to planar CMOS process. The improved lithography in a FinFET technology has the potential to make even very small ETFs very accurate. This is the natural evolution of the work presented in this thesis, and neither ETFs or TD sensors have not been demonstrated in FinFET or other non-planar CMOS technologies.

References

[1] K. R. et. al., "Broadband phonon mean free path contributions to thermal conductivity measured using frequency domain thermoreflectance," *Nature Communications*, vol. 4, no. 1640, 2013.

[2] P. Sverdrup, S. Sinha, M. Asheghi, S. Uma, and K. E. Goodson, "Measurement of ballistic phonon conduction near hotspots in silicon," *Applied Physics Letters*, vol. 78, no. 21, pp. 3331–3333, 2001.

[3] J. A. J. et. al., "Direct Measurement of Room-Temperature Nondiffusive Thermal Transport Over Micron Distances in a Silicon Membrane," *Phys. Rev. Lett.*, vol. 110, p. 025901, Jan 2013.

[4] U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "1650 μm2 thermal-diffusivity sensors with inaccuracies down to $\pm$0.75 ^{o}C in 40nm CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Jan 2016, pp. 206–207.

[5] U. Sonmez, R. Quan, F. Sebastiano, and K. Makinwa, "A 0.008-mm2 area-optimized thermal-diffusivity-based temperature sensor in 160-nm cmos for SoC thermal monitoring," in *ESSCIRC*, Sept 2014, pp. 395–398.

[6] R. Quan, U. Sonmez, F. Sebastiano, and K. A. A. Makinwa, "A 4600 μm2 1.5 ^{o}C (3 σ) 0.9kS/s thermal-diffusivity temperature sensor with VCO-based readout," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2015, pp. 1–3.

[7] U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "Analysis and Design of VCO-Based Phase-Domain $\sigma\delta$ Modulators," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 5, pp. 1075–1084, May 2017.

[8] R. J. Walker, J. A. Richardson, and R. K. Henderson, "A 128x96 pixel event-driven phase-domain $\delta\sigma$-based fully digital 3D camera in 0.13 μm CMOS imaging technology," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2011, pp. 410–412.

[9] Y. H. Liu, A. Ba, J. H. C. van den Heuvel, K. Philips, G. Dolmans, and H. de Groot, "A 1.2 nJ/bit 2.4 GHz Receiver With a Sliding-IF Phase-to-Digital Converter for Wireless Personal/Body Area Networks," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 12, pp. 3005–3017, Dec 2014.

[10] S. M. Kashmiri, K. Souri, and K. A. A. Makinwa, "A Scaled Thermal-Diffusivity-Based 16 Mhz Frequency Reference in 0.16 μm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 7, pp. 1535–1545, July 2012.

[11] S. Pan, Y. Luo, S. H. Shalmany, and K. A. A. Makinwa, "A Resistor-Based Temperature Sensor With a 0.13 pj.k^2 Resolution FoM," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 164–173, January 2018.

A

Appendix

A.1. A Numerical Model for Ballistic Transport in Silicon

As mentioned in Chapter 2, thermal diffusivity of silicon for short distances can be modeled numerically, from the data generated in [1]. The cited experiments had been done via laser grating and the thermal decay times of heat pulses in silicon are measured as a function of laser-grating wave-vector magnitude, defined as *q*. The paper reports the thermal decay rate, defined as γ, with respect to q^2. When thermal diffusion equation is solved under the circumstances of the experiment [1], it can be shown that the decay rate γ should be directly related to *q*:

$$\gamma = \alpha q^2 \tag{A.1}$$

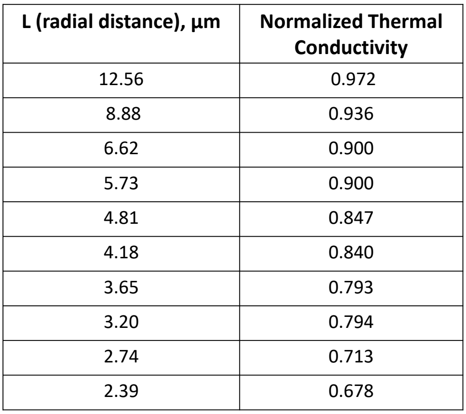
Here, α is thermal diffusivity of silicon and *q* can be directly translated to radial distance *L*:

$$q = 2\pi / L \tag{A.2}$$

The plots presented in the paper can be numerically analyzed to yield a table of normalized thermal diffusivity vs radial distance *L*, as presented in Table A.1. Normalization was done based on the diffusivity of bulk silicon (0.88 cm^2/s).

The dataset is noisy, and does not extend below a radial distance of 2.4μm. It is quite plausible that it was difficult to build silicon structures that can achieve laser-grating less than 2.4μm, and hence no data was reported in the paper for such small distances.

However, the data can be fitted to a suitable function and extrapolated down to 1-2 μm range to get an estimate for ETF behavior at 2μm. The gaussian error function (also known as the erf function) was found to be satisfactory to achieve this



L (radial distance), µm	Normalized Thermal Conductivity
12.56	0.972
8.88	0.936
6.62	0.900
5.73	0.900
4.81	0.847
4.18	0.840
3.65	0.793
3.20	0.794
2.74	0.713
2.39	0.678

Figure A.1: Tabulated data of normalized thermal diffusivity vs distance

fit, and hence the following MATLAB function has been used to generate the normalized model of α with respect to distance used in Chapter 2, Figure 2.2. Further effort would be necessary to validate this mathematical fit at smaller distances.

```
function D_factor = ballistic_model(Lin)
q2 = [0.25 0.5 0.9 1.2 1.7 2.25 2.95 3.85 5.25 6.9];
gamma = [0.0135 0.026 0.045 0.06 0.08 0.105 0.13 0.17 0.208 0.26];
gamma_fit = q2.*(0.25./4.5);
D = gamma./q2.*1e3;
L = 2.*pi./sqrt(q2);
gamma_loss = gamma./gamma_fit;
coef = polyfit(q2,gamma_loss,3);
gamma_fittype = fittype(’0.45+erf(a*10*log10(x)+c)*0.5’);
L_val = Lin*1e6;
q2_val = (2.*pi./L_val)^2;
gamma_loss_f = fit(transpose(L),transpose(gamma_loss),gamma_fittype);
D_factor = gamma_loss_f(L_val);
end
```

References

[1] J. A. J. et. al., “Direct Measurement of Room-Temperature Nondiffusive Thermal Transport Over Micron Distances in a Silicon Membrane,” *Phys. Rev. Lett.*, vol. 110, p. 025901, Jan 2013.

B

Summary

Today's systems-on-chip (SOCs) and microprocessors are complex systems that require multiple temperature sensors to monitor temperature variations in multiple spots on a single silicon die. For such thermal management applications, specialized compact and fast temperature sensors are required. This is necessary because executing an intensive process on an SoC can cause local hotspots in a short amount of time, which can compromise reliability. Such temperature sensors should also be compatible with advanced nanometer CMOS technologies, since complex SoCs and microprocessors are typically implemented in aggressively scaled CMOS processes.

In Chapter 1, the specifications of the temperature sensors required for thermal management are discussed. These requirements can be broken down to five items: area, speed, resolution, accuracy, and power supply compatibility. Compact sensors with <5000 μm^2 area are required, since this allows the sensors to be placed close to hot-spots. Sampling speeds above 1 kSa/s and resolution better than 0.5 ^{o}C are necessary such that the SoC thermal management can respond fast to thermal transients. A 3σ inaccuracy of less than 1 ^{o}C is desired to minimize the margin on SoC's throttle temperature, which translates to better power efficiency at the system level. In order to minimize design and area overhead, the sensors should be powered from the digital supply, which is noisy and can vary significantly (0.6-1.2V). Therefore, sub-1V operation as well as good AC and DC power-supply rejection ratio (PSRR) are necessary.

After this discussion on specifications, a brief overview of CMOS temperature sensing circuits is given with specific focus on the thermal-management requirements. BJT-based temperature sensors are typically the most accurate and achieve good temperature resolution for the given power budget, but usually require large area (>5000 μm^2) and a relatively high supply voltage (>1V), which makes them unsuitable for the sub-1-V nanometer processes used for modern SoCs. MOS-based temperature sensors can be small (<2000 μm^2) and achieve resolution similar to BJT-based sensors, but they are much more inaccurate (more than 2 ^{o}C for one-temperature calibration). Resistor-based sensors can achieve the best temperature resolution, but require extensive calibration at a minimum of two different temperatures to correct their inherent non-linearity, thus resulting in excessive costs. Thermal-diffusivity (TD) based temperature sensors, also known as TD sensors, can be very small, achieve good enough accuracy even without any calibration, and they can operate with sub-1-V supplies. This comes at the cost of worse temperature resolution and energy efficiency, which can be both tolerated in typical thermal management applications.

To tackle this energy efficiency limitation, a different approach compared to previous work must be adopted for the design of Electro-Thermal Filters (ETF), which are the basic component of TD sensors. In an ETF, an electrical signal drives a heating element to produce a heat wave that diffuses in the silicon substrate. The heat is detected by nearby sensors, an, by measuring the time required for the heat to travel from the heater to the detector, the silicon thermal diffusivity can be measured. Since the silicon thermal diffusivity is strongly temperature dependent but well defined, it can be used as temperature sensing principle. Chapter 2 describes the design of a compact energy-efficient ETF. First, the theory of heat diffusion in silicon is treated, including its limitations for heat transport in silicon over μm distances. This theoretical framework is used for understanding the operation of compact ETFs where heat travels only a few μm in silicon and hence exhibits quasi-ballistic (rather than diffusive) transport properties.

Then, three different ETF geometries are introduced: the bar, phase-contour and polygon ETFs. The bar ETF is a simple structure with a long heater implemented by a diffused resistor and the detectors implemented by thermocouples that are orthogonally aligned to the heater. It is easy to implement but it shows low energy efficiency since the large heater causes significant loss of the heat into the silicon substrate. Phase contour ETFs use a small, point heater and the thermocouples' hot junctions are aligned on a phase contour around this heater. This means that all the heat generated by the heater is captured by the thermocouples with the same phase, thus adding constructively. However, this structure suffers from large thermal noise since the thermocouples are typically implemented as long and narrow resistors. The polygon ETF solves this problem by optimizing the thermocouple layout. By maximizing the thermopile's area their resistance is minimized, hence improving signal-to-noise ratio (SNR) of the ETF and the temperature resolution of the TD sensor. This comes at the cost of additional parasitic capacitance, which causes additional phase and consequently temperature inaccuracy.

Despite the drawback in degraded accuracy, the polygon ETF is chosen thanks to its benefits in terms of resolution. Given the polygon ETF geometry, and the theoretical framework of quasi-ballistic heat transport, a model of the ETF behavior is then built to analyze and optimize the ETF design methodology. This model is based on previous work that analyzes the ETF as a complex thermal impedance between the heater and the thermocouples. This model is expanded by further analysis on most significant causes for ETF inaccuracy in silicon: lithography errors, self-heating and mechanical stress. It is expected that lithography errors get smaller with process scaling, although this scaling might not be as aggressive and beneficial to the ETF as once believed. Self-heating is shown to be a significant but non-dominant factor for compact ETFs. It is also shown that mechanical stress is expected to create an error proportional-to-absolute-temperature (PTAT), which can be corrected with a single-point calibration. Armed with this knowledge and the thermal impedance model, two compact polygon ETFs with a radial distance (s) of 2 μm and 3.3 μm have been designed and analyzed. The model's predictions are compared to measurement results obtained in the experiments described in Chapters 4 and 5. Good agreement between the models and silicon results is reached.

Chapter 3 covers the system-level design of phase-domain readouts that convert ETF phase shift into the digital domain. The chapter starts with a brief discussion on phase-detection in CMOS circuits, including coherent demodulation, and then expands to Phase-domain ΣΔ modulator (PDΣΔM) architecture. Here, the goal is to show that the PDΣΔM is one of the simplest and most efficient architectures for ETF readout. This simplicity translates into a small area of the circuit implementation. A PDΣΔM based on a Gm-C front-end is first considered. In order to reduce the size of the integration capacitor to meet the application area requirement, the two-step conversion technique is introduced. By running a short coarse conversion, the ETF's phase shift can be digitized with an error below 5^o. Then, a longer, fine conversion can be run with a much smaller range, which is chosen based on the coarse conversion result. This reduction by a factor 8x in the required ΣΔ modulator range significantly reduces the swing on the integration capacitor. For a given supply-limited voltage swing, this allows for a smaller capacitor and hence a more compact sensor.

However, it is shown that the PDΣΔM with two-step conversion is still too big in nanometer CMOS, and does not benefit much from CMOS scaling. Therefore, an area-efficient and more digital-friendly architecture based on a voltage-controlled oscillator (VCO) and an up/down counter is discussed. This architecture, called the VCO-based $PD\Sigma\Delta M$, achieves coherent demodulation by converting the ETF's voltage signal into the frequency domain by means of a VCO and then using an up/down counter to demodulate the ETF's phase shift. It is shown that it behaves similar to an ideal Gm-C based PDΣΔM with additional quantization noise caused by the finite sampling operation of the up/down counter. In further analysis of this additional noise source, it is shown that a VCO with large gain (K_{VCO}) is required to suppress its quantization noise. It is also shown that an 8 bit counter is usually sufficient for correct operation, and that the counter can be allowed to 'wrap around', i.e. to overflow, during operation. Since the up/down counter allows effortless multi-bit sampling at its output, a multi-bit PDΣΔM, rather than a two-step conversion, is implemented with this architecture. The design trade-offs of this architecture are discussed in detail in the rest of Chapter 3. In particular, it is shown that a Gm-C or VCO-based PDΣΔM is robust to non-linearity in the amplitude domain, as long as two-step conversion or multi-bit PDΣΔMs are used. This robustness to non-linearity is critical in achieving a compact design with acceptable inaccuracy, since it relaxes the VCO's linearity specification. It also allows high-density MOS capacitors to be used as integration capacitor element for Gm-C based architecture.

In chapter 4, the detailed implementation of a TD-based temperature sensor in

160-nm CMOS is presented. This prototype was used to evaluate the behavior of compact ETFs and the area-efficient Gm-C PDΣΔM readout. Circuit techniques, such as single-shot auto-zeroing to eliminate Gm-stage offset, and heater-drive inversion (HDI) to suppress the impact of electrical cross-talk in the ETF, are introduced. The chapter also discusses the design of the energy-efficient telescopic Gm-stage and the heater drive logic in detail. Occupying only 8000 μm^2 area and consuming 3 mW from a 1.8 V supply, the prototype design achieves a resolution of 0.21 °C within 1-ms conversion time. The untrimmed 3σ inaccuracy of the sensors is 2.4 °C, which improves to 0.65 °C after a single-temperature trim. This design is intended to be the first stepping stone towards a scaled design in nanometer CMOS.

Since the prototype design is so small, it has been laid out as an array of temperature sensors, which allows us to replicate and validate the thermal management problem in modern SoCs. The response of 6 TD-based temperature sensors to a step heat pulse, which is generated by an on-chip test heater is shown. It is demonstrated that the sensor closest to the test heater is subjected to higher temperature (~7 °C) with a sharp gradient (~1 °C/ms) compared to sensors that are far away (>100μm), which only observe a temperature increase of (~4 °C) with a slower time constant (~0.4 °C/ms). This experiment demonstrates why it is important for a complex SoC to have multiple temperature sensors over the same silicon die.

Finally, chapter 5 presents the implementation of a TD-based temperature sensor in 40-nm CMOS, occupying only 1650 μm^2 area and consuming 2.5 mW from 0.9-1.2-V supply. The design is also the first implementation of ETFs and VCO-based PDΣΔMs in 40-nm CMOS, and the first sub-1V TD-based temperature sensor. Moreover, a foreground calibration technique that improves the readout's accuracy is demonstrated. Such technique is based on measuring the phase error of the ETF (due to parasitics) and the readout in the electrical rather than in the thermal domain, and compensating for this error during normal operation. Since the ETF's electrical phase error is calibrated by this so-called phase-calibration technique, it is well-suited to be used with polygon ETFs which normally suffer from this problem. Furthermore, it is shown that a compact 3-inverter ring current-controlled oscillator (CCO) combined with a Gm-stage is sufficient to meet the K_{VCO} and noise requirements. It consumes less than 0.14 mA current from a 0.9-1.2-V supply and occupies less than 500μm^2 area. Due to the relaxed linearity requirements of the multi-bit PDΣΔM, a VCO second-order non-linearity of only -55 dB is considered acceptable for this design. A post-VCO amplifier is used to level shift the VCO output up to digital levels suitable for the operation of the cascaded counter. Such amplifier consumes only 50 μA and provides robust level shifting from the variable VCO output amplitude to the counter. The up/down counter consumes only 0.3 mA current at 500 MHz input frequency and is implemented by standard digital place-and-route (PnR) logic, which is area efficient and demonstrates that a custom design is not necessary.

Similar to the 160-nm CMOS design, multiple TD sensors have been laid out in an array fashion. These include two scaled ETFs that can achieve down to 0.26 °C resolution within 1-ms conversion time, and a room-temperature-trimmed 3σ inaccuracy of 0.65 ^{o}C. Further measurements show that plastic-packaged ETFs demonstrate higher inaccuracy. However, as predicted before in Chapter 2, it is shown that a room temperature PTAT calibration eliminates this additional inaccuracy. After such calibration, the 3σ inaccuracy is 0.7 ^{o}C for plastic packaged TD sensors.

Chapter 6 presents the conclusions of this thesis. A summary is made of its novel contributions: an improved ETF design, architectural improvements to $PD\Sigma\Delta M$s and circuit implementations of such architectures. A section on future work discusses possible improvements in future designs. It is expected that ETF design and TD sensors would have further applications in thermal management and heat transport experiments. Moreover, since the implemented TD sensors are very small and accurate, they can be used to calibrate higher resolution but more inaccurate temperature sensors such as those based on MOSFET or resistors.

C
Samenvatting

De huidige systems-on-chip (SOC's) en microprocessors zijn complexe systemen met meerdere temperatuursensoren om temperatuurvariaties op meerdere plekken op de chip te controleren. Voor deze toepassingen zijn gespecialiseerde kleine en snelle temperatuursensoren nodig. Dit is nodig omdat het uitvoeren van een intensief proces op een SoC hotspots kan creëren, die de betrouwbaarheid in gevaar kunnen brengen. Dergelijke temperatuursensoren moeten ook compatibel zijn met geavanceerde CMOS-nanometertechnologieën.

In hoofdstuk 1 worden de specificaties van de temperatuursensoren die nodig zijn voor deze toepassing voor thermisch beheer besproken. Deze specificaties zijn: ruimte, snelheid, resolutie, nauwkeurigheid en stroomvoorziening. Compacte sensoren met <5000μm ^{2} ruimte zijn vereist, omdat hierdoor de sensoren in de buurt van hotspots kunnen worden geplaatst. Snelheden hoger dan 1 kSa / s en een resolutie beter dan 0.5^{o}C zijn noodzakelijk zodat het SoC snel kan reageren op thermische transiënten. Een 3σ onnauwkeurigheid minder dan 1 ^{o}C is ook gewenst. De sensoren moeten worden gevoed door de digitale stroomvoorziening, die heeft veel ruis is en kan variëren (0.6-1.2V).

Vervolgens wordt een kort overzicht gegeven van CMOS-temperatuursensoren met specifieke focus op de vereisten voor thermisch beheer. BJT-temperatuursensoren zijn de beste nauwkeurige en bereiken een goede resolutie, maar nemen een groot oppervlak (>5000 μm^2) en >1V stroomvoorziening. Daaroom zijn ze ongeschikt voor de sub-1V nanometerprocessen die worden gebruikt voor moderne SoC's. MOS-temperatuursensoren kunnen klein zijn (<2000μm^2) en een goede resolutie bereiken, maar ze zijn onnauwkeuriger (meer dan 2 ^{o}C voor kalibratie op één temperatuur). Weerstand-temperatuursensoren heeft de beste temperatuurresolutie, maar vereisen kalibratie bij minimaal twee verschillende temperaturen. Dit resulteert in ekstra kosten. Thermische diffusiviteit (TD) temperatuursensoren, ook bekend als TD-sensoren, kunnen erg klein zijn, zelfs zonder kalibratie een voldoende goede nauwkeurigheid bereiken, en ze kunnen werken met sub-1V stroomvoorziening. Maar ze hebben slecht temperatuur resolutie en energie-efficiëntie, die beide kunnen worden getolereerd in typische warmtebeheertoepassingen.

Om deze energie-efficiëntiebeperking op te lossen, wordt een andere methode gebruikt voor het ontwerp van elektro-thermische filters (ETF), die de basiscomponent van TD-sensoren zijn. In een ETF zit een verwarmingselement en een paar warmtesensoren. Een elektrisch signaal stuurt het verwarmingselement aan, dat een warmtesignaal produceert. Deze diffundeert in het siliciumsubstraat en neemt de sensoren na enige tijd op. Aangezien de thermische diffusiviteit van silicium sterk temperatuurafhankelijk maar goed gedefinieerd is, kan het ETF worden gebruikt als een temperatuursensor. Hoofdstuk 2 beschrijft het ontwerp van een kleine energie-efficiënte ETF. Ten eerste wordt de theorie van warmtediffusie in silicium behandeld, inclusief de beperkingen voor warmtetransport in silicium over afstanden van een paar μm.

Vervolgens worden drie verschillende ETF-geometrieën geïntroduceerd: de staaf, fase-contour en veelhoek ETF's. De staaf-ETF is een eenvoudige structuur met een lange verwarmer geïmplementeerd door een diffuse weerstand en de detectoren geïmplementeerd door thermokoppels die orthogonaal zijn uitgelijnd met de verwarmer. Het is eenvoudig te implementeren, maar het heeft een lage energie-efficiëntie omdat de grote verwarmer warmte verliest in het siliciumsubstraat. Fasecontour ETF's gebruiken een kleine, puntverwarmer en de hete knooppunten van de thermokoppels zijn uitgelijnd op een fasecontour rond deze verwarmer. Dit betekent dat alle warmte die door de verwarmer wordt gegenereerd, wordt opgevangen door de thermokoppels. Deze structuur heeft echter veel ruis, omdat de thermokoppels lange en smalle weerstanden zijn. De veelhoek ETF lost dit probleem op door de layout van het thermokoppel te optimaliseren. De weerstand van de thermozuil is geminimaliseerd, waardoor de signaal-ruisverhouding (SNR) van de ETF en de temperatuurresolutie van de TD-sensor wordt verbeterd. Dit gaat ten koste van extra parasitaire capaciteit, die de fase- en temperatuuronnauwkeurigheid verslechtert.

Ondanks het nadeel van verslechterde nauwkeurigheid, wordt de veelhoek ETF gekozen dankzij de goede resolutie. Vervolgens wordt een model van de veelhoek ETF gebouwd om de ETF te optimaliseren. Dit model is gebaseerd op eerder werk dat de ETF analyseert als een complexe thermische impedantie tussen de verwarmer en de thermokoppels. Dit model wordt uitgebreid door verdere analyse van de belangrijkste oorzaken van ETF-onnauwkeurigheid in silicium: lithografiefouten, zelfverhitting en mechanische spanning. Verwacht wordt dat lithografiefouten kleiner worden bij processchaling, hoewel deze schaalverdeling mogelijk niet zo voordelig is als ooit werd gedacht. Zelfverwarming is een significante factor voor compacte ETF's. Mechanische spanning resulteert in fouten proportioneel-tot-absolute-temperatuur (PTAT), die kunnen worden gecorrigeerd met een temperatuurkalibratie. Vervolgens zijn twee compacte veelhoek ETF's met radiale afstanden (s) van 2μm en 3.3μm ontworpen en geanalyseerd. Goede overeenstemming tussen de modellen en experimentresultaten wordt bereikt.

Hoofdstuk 3 behandelt het ontwerp op systeemniveau van fase-domeinomzetters die de ETF-fase in digitaal omzetten. Het hoofdstuk begint met een korte discussie over fasedetectie in CMOS-circuits, inclusief coherente demodulatie, en legt vervolgens de architectuur van het fase-domein ΣΔ modulator (PDΣΔM) uit. EenPDΣΔM op basis van een Gm-C front-end wordt eerst beschreven. Om de grootte van de integratiecondensator te verminderen, wordt de tweestaps-conversie techniek geïntroduceerd. Door een korte conversie uit te voeren, kan de faseverschuiving van de ETF worden gedigitaliseerd met een fout van minder dan 5^o. Vervolgens kan een langere, fijne conversie worden uitgevoerd met een veel kleiner bereik, dat wordt gekozen op basis van het korte conversie-resultaat.Deze reductie met een factor 8x in het modulatorbereik vermindert de swing op de integratiecondensator. Dit zorgt voor een kleinere condensator en dus een kleinere sensor.

De PDΣΔM met tweestaps-conversie techniek is echter nog steeds te groot in nanometer CMOS en profiteert niet veel van CMOS-schaal. Daarom is een nieuwe architectuur geïntroduceerd met een voltage-controlled-oscillator (VCO) en een omhoog/omlaag digitaalteller. Deze zogenaamde op VCO gebaseerde PDΣΔM zet het ETF signaal om in het frequentiedomein en demoduleert dit met een digitaalteller. Het is vergelijkbaar met een Gm-C PDΣΔM met extra kwantisatie-ruis veroorzaakt door de omhoog/omlaag digitaalteller. Er is aangetoond dat een VCO met grote versterking (K_{VCO}) deze kwantisatieruis kan verminderen. Er wordt ook aangetoond dat een 8-bits teller goed is voor een correcte werking en dat de teller kan 'overlopen'. De ontwerpdetails van deze architectuur worden in detail besproken in de rest van hoofdstuk 3. In het bijzonder is aangetoond dat een op Gm-C of VCO PDΣΔM robuust tot niet-lineariteit is, zolang tweestaps-conversie of een multi-bit modulator worden gebruikt. Dit ontspant de lineariteitsspecificatie van de VCO en maakt het sensor klein.

In hoofdstuk 4 wordt de implementatie van een TD-sensor in 160-nm CMOS gepresenteerd. Dit prototype werd gebruikt om het gedrag van compacte ETF's en de Gm-C PDΣΔM te evalueren. Technieken, zoals single-shot auto-zeroing om Gm-stage offset te elimineren, en heater-drive inversion (HDI) om de impact van elektrische overspraak in de ETF te verminderen, worden geïntroduceerd. Het hoofdstuk bespreekt ook het ontwerp van de energiezuinige telescopische versterker en de logica van de verwarmingsaandrijving. Met een ruimte van slechts 8000μm^2 en een verbruik van 3mW uit een 1.8V stroomvoorziening, bereikt het prototypeontwerp een resolutie van 0.21°C binnen een conversietijd van 1ms. De 3σ onnauwkeurigheid van de sensoren is 2.4°C, wat verbetert tot 0.65°C na een trim van een temperatuur. Dit ontwerp is bedoeld als eerste opstap naar een geschaald ontwerp in nanometer CMOS.

Omdat het prototypeontwerp zo klein is, zit de sensoren in een reeks, waarmee we het probleem van het thermische beheer in moderne SoC's kunnen repliceren en valideren. Er is aangetoond dat de sensor die zich het dichtst bij de testverwarming bevindt, een hogere temperatuur (~7°C) heeft met een scherpe thermische gradiënt (~1°C/ms). Sensoren die ver weg zijn (>100μm) nemen alleen een temperatuurstijging van ~ 4°C) waar met een langzamere tijdconstante (~0.4°C/ms).

Hoofdstuk 5 presenteert de implementatie van een TD-sensor in 40-nm CMOS, die slechts 1650μm^2 ruimte neemt en 2.5mW verbruikt van 0.9 tot 1.2V stroomvoorziening. Dit ontwerp is ook de eerste implementatie van ETF's en VCO PD ΣΔM in 40-nm CMOS, en de eerste sub-1V TD-temperatuursensor. Er wordt een voorgrondkalibratietechniek aangetoond die de nauwkeurigheid van de uitlezing verbetert. Deze techniek is gebaseerd op het meten van de fasefout van de ETF (vanwege parasieten) en de uitlezing in het elektrische in plaats van in het thermische domein, en het compenseren van deze fout tijdens normaal operatie. Verder is aangetoond dat een compacte 3-inverter oscillator (CCO) gecombineerd met een versterker goed is om te voldoen aan de specificaties van K_{VCO} en ruis. De gehele front-end verbruikt minder dan 0.14mA stroom en neemt minder dan 500μm^2 ruimte. Vanwege de ontspannen lineariteitsspecificaties van de multi-bit PDΣΔM, is een VCO tweede-orde niet-lineariteit van slechts -55dB goed. Een post-VCO-versterker wordt gebruikt om de VCO-signaal naar een niveau te verschuiven dat geschikt is voor de werking van de omhoog/omlaag digitaalteller. De digitaalteller verbruikt slechts 0.3mA stroom bij een frequentie van 500MHz en wordt geïmplementeerd door standaard digitale plaats-en-route(PnR) logica.

Vergelijkbaar met het 160-nm CMOS-ontwerp, zijn meerdere TD-sensoren op een reeks-manier geplaats. Deze omvatten twee geschaalde ETF's die een resolutie tot 0,26°C kunnen bereiken binnen een conversietijd van 1 ms en een onnauwkeurigheid van 0.65^{o}C. Verdere metingen tonen aan dat ETF's met een plastic verpakking een hogere onnauwkeurigheid hebben. Zoals eerder in hoofdstuk 2 werd voorspeld, elimineert een PTAT-kalibratie deze extra onnauwkeurigheid. Na deze PTAT kalibratie is de 3σ onnauwkeurigheid 0.7^{o}C voor in plastic verpakte TD-sensoren.

Hoofdstuk 6 presenteert de conclusies van dit proefschrift. Een samenvatting wordt gemaakt van zijn nieuwe bijdragen: een verbeterde ETF, architecturale verbeteringen aan PDΣΔMs en circuit technieken. Een sectie over toekomstig werk bespreekt mogelijke verbeteringen in toekomstige ontwerpen. Verwacht wordt dat ETF-ontwerp en TD-sensoren verdere toepassingen zouden hebben bij experimenten met thermomanagement en warmtetransport. Omdat de geïmplementeerde TD-sensoren erg klein en nauwkeurig zijn, kunnen ze bovendien worden gebruikt voor het kalibreren van hogere resolutie maar meer onnauwkeurige temperatuursensoren zoals die op MOSFET of weerstanden zijn gebaseerd.

Acknowledgements

It is said that it takes a village to raise a child. During my experience as a PhD student, I have learned it takes an even bigger village to graduate from PhD. In this section, I would like to thank everyone in this special 'village', who helped me guide through the process.

At the forefront are my two promotors, Fabio Sebastiano and Kofi Makinwa. They have supported and pushed me through the darkest of times, through many technical and scheduling challenges. Their technical ingenuity, perseverance and patience are what makes them excellent academics and teachers. Even though they can be sometimes tough on their students, I have seen they always stand behind them in their hour of need. Fabio has been there for me even in my most frustrated moments, and helped me get on track with his support and excellent people skills. It has been a great relief to communicate my concerns and frustrations on him, which he always took graciously. Looking back, I thoroughly enjoyed my journey with them. I would also like to thank STW for fully funding this research, all the way to its successful completion.

Prof. Huisjing and Bult have also helped me grow, during my time as a teaching assistant in EI lab courses. Thank you for the opportunity, and the great coaching. As the faculty secretary, Joyce Siemers has been of great help. I am sure I would not have survived the bureacuracy otherwise! I owe a great debt to our technicians Zu Yao, Lukasz Pakula and Ron van Puffelen who have kept our labs alive even under moments of extreme stress.

I would like to thank my ex-colleague and friend Burak Gonen, whom I shared a lot of PhD-life experiences and had many technical discussions over the years. His support was vital throughout this process. My roommate Nishant Lawand has been also there for me, and has given me important tips to get my life started properly in the Netherlands.

My current colleagues Kamran Souri, Kia Souri, and Saleh H. Shalmany also deserve a special spot. We have shared many evenings in the oven room, which would have been otherwise unbearable. It is a great joy to be able to keep working with this energetic team for the last 8 years! Kamran and Saleh have also helped me revise this thesis, which has greatly helped with the quality of English.

During my time as a PhD candidate, I was also involved in daily supervision of several master students. I would like to thank them specifically because I have learned a lot during this process. I would like to thank Rui Quan for his design, layout and experimental work in the first VCO-based PDSDM, Jan Angevare for many technical discussions and experimental support, and Sining Pan for this theoretical analyses and always pushing the boundaries of what is possible.

I would like to also thank my other colleagues Lorenzo Pedala, Cagri Gurleyuk, Qinwen Fan, Pierluigi Cenci, Junfeng Jiang, N. Pelin Ayerden and Zhichao Tan, for all of our discussions and many life experiences we shared together. I would like to thank Thije Rooijers for his help on his help with the Dutch translation.

Last of all, I would like to thank my family for their monumental support, care and encouragement. A special mention within this group goes to Tara, whose love and care during many long ’thesis nights’ are much appreciated. She has been my support pillar, and her love has ensured that I was able to wrap up this process.

List of Publications

Journal Papers

1. **U. Sonmez**, F. Sebastiano, K.A.A. Makinwa, 'Compact Thermal-Diffusivity-Based Temperature Sensors in 40-nm CMOS for SoC Thermal Monitoring', *IEEE Journal of Solid State Circuits*, vol. 52, no. 3, pp. 834-843, March 2017.

2. **U. Sonmez**, F. Sebastiano, K.A.A. Makinwa, 'Analysis and Design of VCO-Based Phase-Domain $\Sigma\Delta$ Modulators', *IEEE Journal of Solid State Circuits*, vol. 64, no. 5, pp. 1075-1084, May 2017.

3. **U. Sonmez**, H. Kulah, T. Akin, 'A $\Sigma\Delta$ micro accelerometer with 6 μg$\sqrt{Hz}$ resolution and 130dB dynamic range', *Analog Integrated Circuits and Signal Processing*, vol. 81, no. 2, pp. 471-485, Aug 2014.

Conference Papers

1. **U. Sonmez**, R. Quan, F. Sebastiano, K.A.A. Makinwa, 'A 0.008-mm^2 area-optimized thermal-diffusivity-based temperature sensor in 160-nm CMOS for SoC thermal monitoring', *ESSCIRC 2014*, pp. 395-398, Sept 2014.

2. R. Quan, **U. Sonmez**, F. Sebastiano, K.A.A. Makinwa, 'A 4600 μm^2 1.5 ^{o}C (3σ) 0.9 kS/s Thermal-Diffusivity Temperature Sensor with VCO-based Readout', *ISSCC Dig. of Tech. Papers*, Feb 2015.

3. B. Yousefzadeh, **U. Sonmez**, N. Mehta, J. Borremans, M.A.P. Pertijs, K.A.A. Makinwa, 'A generic read-out circuit for resistive transducers', *IWASI*, pp. 122-125, June 2015.

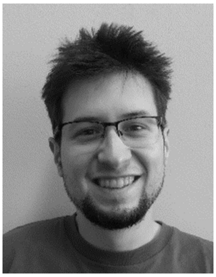
4. J. Angevare, L. Pedala, **U. Sonmez**, F. Sebastiano, K.A.A. Makinwa, 'A 2800-μm^2 thermal-diffusivity temperature sensor with VCO-based readout in 160-nm CMOS', *Asian Solid-State Circuits Conference*, Nov 2015.

5. **U. Sonmez**, F. Sebastiano, K.A.A. Makinwa, 'A 1650μm^2 Thermal-Diffusivity Sensors with inaccuracies down to $\pm$0.75^{o}C in 40nm CMOS, *ISSCC Dig. of Tech. Papers*, pp. 395-398, Feb 2016.

6. L. Pedala, **U. Sonmez**, F. Sebastiano, K.A.A. Makinwa, K. Nagaraj, 'An oxide electrothermal filter in standard CMOS', *IEEE Sensors*, Nov 2016.

7. **U. Sonmez**, H. Kulah, T. Akin, 'A fourth order unconstrained $\Sigma\Delta$ capacitive accelerometer', *2011 International Solid-State Sensors, Actuators and Microsystems Conference*, pp. 707-710, June 2011.

About the Author



Uğur Sönmez (S'10-M'15) was born in Istanbul, Turkey on 3 April 1986. He obtained his B.Sc. and M.Sc. degrees in electronics from Middle East Technical University, Ankara, Turkey in 2008 and 2011, respectively. From 2011 to 2016, he pursued his Ph.D. at Electronic Instrumentation Laboratory in TU Delft.

He joined SiTime in 2016, and has been subsequently working as an analog/mixed-signal IC designer. His current research interests include sensor interfaces, precision analog circuits, data converters, oscillators and PLLs.