A Highly-Scalable Thermal-Diffusivity-Based Temperature Sensor

J. A. Angevare

Electronic Instrumentation Laboratory

TU Delft Delft University of Technology

Electronic Instrumentation Laboratory

Challenge the future

A Highly-Scalable Thermal-Diffusivity-Based Temperature Sensor

in 160 nm CMOS for thermal monitoring applications

by

J. A. Angevare

in partial fulfillment of the requirements for the degree of

Master of Science
in Microelectronics

at the Electronic Instrumentation Laboratory,
Department of Electrical Engineering,
Delft University of Technology,
to be defended publicly on Thursday December 17, 2013 at 3:00 PM.

Supervisor:	Prof. dr. K. A. A. Makinwa
Thesis committee:	Prof. dr. K. A. A. Makinwa, TU Delft
	Dr. F. Sebastiano, TU Delft
	Prof. dr. Ing. L. de Vreede, TU Delft

This thesis is confidential and cannot be made public until December 31, 2020.

TU Delft Delft University of Technology

Abstract

A highly digital temperature sensor, based on the thermal diffusivity of silicon, for thermal management applications has been reported. Recently, Thermal-Diffusivity based temperature sensors have undergone a radical change from an analog-based readout architecture to a more digital VCO-based one. However the more digital approach led to significantly worse performance, i.e. accuracy and resolution. The proposed sensor uses the same digital approach but achieves much better performance.

The sensor achieves an inaccuracy of ±2.9 °C (3σ) from -35 °C to 125 °C with no trimming and ±1.2 °C (3σ) after a single-point trim, while achieving a resolution of 0.47 °C (rms) at 1 kSa/s. Its compact area (2800 μm^2) is enabled by the adoption of a VCO-based phase-domain ADC. Since 53% of the sensor area is occupied by digital circuitry, the sensor can be easily ported to more advanced CMOS technologies with further area reduction, which makes it well suited for thermal monitoring in microprocessors and other systems-on-chip.

Acknowledgments

First and foremost I would like to thank Professor Kofi Makinwa for all opportunities, feedback and support given. He is the mastermind behind most ideas and a visionary in executing them. He took me along to San Fransisco and Xiamen, and showed me a good time, not only in San Francisco and Xiamen, but also in Delft. I cannot but have great respect for his ability to come up with new ideas and his passion for his work.

The next person to thank is Uğur Sönmez, my daily supervisor. Every time I've had a problem he came by and helped me. Moreover, I don't think there is a single problem he cannot solve. Also a heart felt thanks to Fabio Sebastiano, who will always make time to help anyone. He has been a great support in reviewing my layout, and has been responsible for writing most of the published conference paper.

I would also like to thank the group at the Electronic Instrumentation Laboratory. I don't think it is ever really apparent what work goes on in the background, nor how important it is for us to be able to do what we do. That's why I want to put in a special thanks to the staff, for helping me with my measurement setup, for maintaining the laboratory equipment and for helping me navigate through all the forms of bureaucracy.

Also a special thanks to the reviewers; Constantino, Junfeng, Lorenzo, Luke, Sining, Uğur, Vincent and Yikun. Also an extra special thanks for Junfeng for being my very thoughtful guide and translator while in China.

Last but not least; a special thanks to my family for their silent support and unwavering trust in me. Without you this entire episode of my life would never even have been possible!

J. A. Angevare
Delft, December 2015

Contents

List of Figures

1

Introduction

The thermal monitoring of multi-core processors is becoming an increasingly hot topic [1]. With the ever shrinking dimensions of CMOS technologies more and more transistors fit in the same surface area. This has been the main force behind the dramatic reduction in the cost of logical functions. However the energy required to switch a digital gate has not reduced as dramatically as the gate dimensions, resulting in rapidly increasing energy densities, and hence die temperatures, of circuits realized in modern processes.

The result of such on-chip energy densities is that local islands of high temperature, i.e. hot-spots, may be created. Depending on the exact algorithm being executed, such hot-spots may occur in different parts of the chip. This problem has been further aggravated by the recent shift towards distributed processing (such as multi-core architectures).

When die temperatures become excessive, say beyond 75 °C, it may result in reliability issues. For instance, the mobility of charge carriers will decrease, resulting in longer delays. Even worse, at high temperatures MOS devices leak more, which in itself consumes power, resulting in further heating. Finally, since dopant diffusion is exponentially dependent on temperature, too much heat may damage or, at least, dramatically reduce the lifetime of a chip.

All of this adds up to the fact that thermal monitoring of high-performance digital circuitry is a necessity. The next section will explain the requirements of thermal monitoring systems and also the actions which can be taken to avert reliability issues or permanent damage due to overheating.

1.1. Thermal Management

In order to understand the requirements for thermal management it is necessary to first understand the actions which can be taken to prevent overheating. There are two groups of techniques: the first involves reducing heat production, while the second attempts to balance it. Under the first group fall stop-go policies as well as dynamic voltage and/or frequency scaling (DVFS) techniques. Stop-go policies simply aim to stop the execution of the circuitry that causes overheating, while DVFS techniques reduce circuit performance and so heat production. The second group of techniques, which balance heat-production, make use of the increasing parallelism in high performance digital circuitry by monitoring local heat production and intelligently scheduling the workload. Obviously the last group requires more digital circuitry and/or software but also has the highest performance. The trend in high performance digital

circuits is towards continuous workload balancing, which in turn drives the need for better thermal monitoring.

Irrespective of how overheat prevention is done, there is a requirement to monitor local hot-spots. Since local hot-spots can rapidly increase the local die temperature (time constants are in the order of milliseconds) there is a need for multiple local temperature sensors with millisecond response times. Because many of these sensors are required, their area should be as small as possible. In short, in order to ensure that local die temperatures remain within a given range, quick, accurate and local sensors are required.

However, most integrated temperature sensors must be calibrated to counteract the effect of process spread. In thermal monitoring applications, the need to calibrate multiple sensors results in a significant increase in production cost. It would therefore be nice to have temperature sensors which are accurate enough "out of the box".

From the above, we can conclude that the ideal qualities of a temperature sensor for thermal monitoring are: small area, good untrimmed accuracy and conversion times in the order of (sub)milliseconds.

1.2. CMOS compatible temperature sensors

For thermal monitoring applications, we need to actively monitor the temperature of the die itself. Therefore the temperature sensitive part of the sensor should be as close to, or rather be in the die. Furthermore, costs would be greatly reduced if the sensor could be made with standard CMOS technologies.

Four kinds of temperature sensors satisfy these conditions: bandgap sensors (made with either BJT's, MOSFET's or diodes), resistor-based sensors, (electrical) delay line elements or thermal diffusivity sensors. A quick review will show which of these is most suited for thermal monitoring applications.

1.2.1. Bandgap

Traditionally, bandgap sensors are the most widely used and come in a variety of flavors. In [2] a PTAT (proportional to absolute temperature) voltage is digitized with respect to a reference voltage, made from a PTAT and CTAT (complementary to absolute temperature) voltage. In the following, any sensor using a PTAT or CTAT voltage created by either a BJT or MOS device will be referred to as a bandgap sensor. In some cases, external reference voltages may be used, e.g. as in [3].

BJT-based bandgap sensors are the most popular, with MOSFET-based sensors slowly gaining popularity, mainly because they require less headroom. Another factor contributing to this is the fact that modern CMOS processes usually do not provide well-optimized BJTs, but only provide so called parasitic PNPs. These devices make use of the same diffusions required to make a PMOS device: The base is made from the n-well, the emitter is made from the p+ drain/source, and the collector is made from the p-type substrate, and so is tied to ground. Since the basic idea behind all bandgap sensors is roughly the same, we will only analyze the more common BJT flavor.

BJT based temperature sensors work by forcing a current through a set of BJT's generating a temperature dependent voltage. This voltage is CTAT by equation (1.1), the difference between

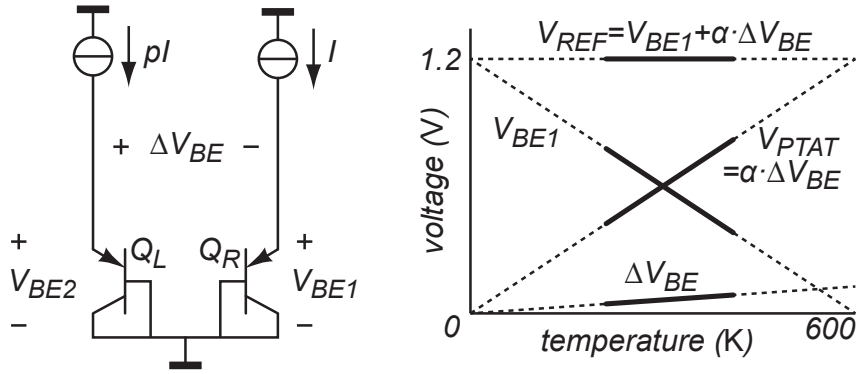

Figure 1.1: Temperature sensing principle of BJT based temperature sensors (remade from [4]).

two of these junctions with different current densities is then (1.2).

$$V_{\mathrm{BE}}(T) = \frac{kT}{q}\ln\left(\frac{I_{\mathrm{bias}}(T)}{I_S(T)}\right) \tag{1.1}$$

$$\Delta V_{\mathrm{BE}}(T) = \frac{kT}{q}\ln(p) \tag{1.2}$$

Adding these two values in a correct ratio a temperature independent voltage (better known as a bandgap voltage) can be generated (see figure 1.1). Finally the $\alpha \cdot \Delta V_{\mathrm{BE}}$ PTAT voltage can be compared against this bandgap voltage to do a temperature reading.

The main downside of bandgap sensors is that they do not benefit from technology scaling, which mainly involves the improvement of lithography. Good BJTs, however, require wll-defined doping profiles. Also newer technologies often don't have 'native' support for BJT's, which must be made as parasitic devices.

Another downside is that BJTs naturally provide a temperature dependent output voltage. With newer technologies, supply voltages decrease at a faster rate than the threshold-voltage, resulting in smaller headroom. Therefore porting a BJT temperature sensor design to a newer technology proves to be far from trivial.

1.2.2. RESISTOR

Resistor based sensors rely on the sensitivity of a resistance to temperature. The temperature sensitive resistor (thermistor) can be used as part of an RC filter, which has the advantage that the readout can now be done in the phase-domain. One such example of a resistive temperature sensor is [5], which uses a Wien-bridge topology. This was (see figure 1.2) used to obtain a reasonably linear phase-shift over a small frequency and temperature range.

However, resistor based temperature sensors are plagued by similar downsides as bandgap sensors. Just like bandgap sensors, resistor based sensors rely heavily on doping for their electrical characteristics, therefore, to minimize spread, they require a larger size. Moreover the temperature dependency of resistors is often very non-linear and spreads from sensor to sensor. Therefore resistive temperature sensors need multi-point trimming to achieve any kind of competitive accuracy; this makes them ill-suited for thermal monitoring applications.

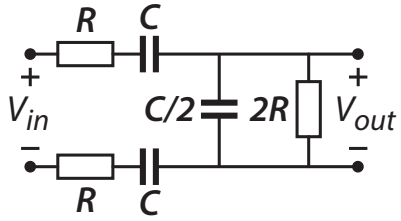

Figure 1.2: Wien Bridge.

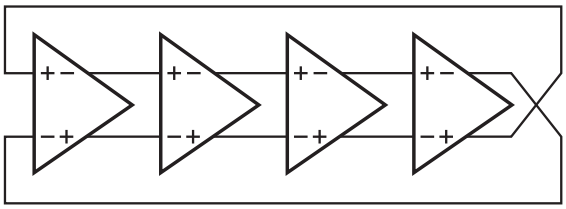

Figure 1.3: Pseudo Differential Ring Oscillator.

1.2.3. Inverter Delay

Another class of temperature sensors is based on inverter (or gate) delay. Usually a ring-oscillator is made, whose oscillation period depends on the threshold-voltage and the charge carrier mobility of MOSFETs, which in turn depends on the temperature. An alternative is to use a TDC to measure the delay of a set of inverters, however TDCs suffer from their own downsides such as large spread, high power consumption and high power supply sensitivity. Even with ideal readout, the temperature dependent delay is still very non-linear and process dependent, therefore temperature sensors based on inverter delay usually need multi-point trimming.

In [6] the authors employ two delay lines, with different biasing, to cancel out this non-linearity and suppress process variation, which is present in both delay lines. Normally the PSRR of inverter delay based sensors is very bad, however the authors claim that, using a pseudo-differential structure (see figure 1.3) and supply regulator, they achieve a good suppression of the supply noise. The requirement for a supply regulator, which is included in the area, makes inverter delay based sensors even less attractive.

1.2.4. Leakage

Leakage based temperature sensors are often used to measure process variation and/or temperature. Modern technologies suffer from multiple sources of leakage; sub-threshold leakage, gate tunneling, punch-through current and reverse bias current of the drain diffusion diode.

In [7], the authors assume sub-threshold leakage to be dominant, and used that leakage current to make an oscillator. A supply voltage range from 0.5 V to 1.0 V was claimed, however the supply sensitivity was not stated, and since the design uses a ring oscillator, its supply rejection will not be very good. The oscillator period was close to the target for a thermal monitoring application, however the noise for such a measurement was not reported. Although the reported accuracy and area for state-of-the-art leakage based temperature sensors are very good, their downsides, such as poor PSRR and resolution, appear to be too bad to be reported. Both PSRR, because of the high-power high-density digital circuitry surrounding the thermal monitoring sensor, and the detection limit are very important for thermal monitoring applications. Therefore leakage based sensors are also not ideal for this application.

1.2.5. Thermal Diffusivity

Thermal-Diffusivity based temperature sensors use the temperature dependency of Thermal-Diffusivity, in other words; the time heat takes to propagate through a material depends on the material's temperature. By generating small heat pulses in the substrate of a chip, a time-delay, based on the thermal-diffusivity of silicon, can be read-out. Because this is in principle a time measurement, TD sensors don't suffer from the shrinking voltage headroom in smaller tech-

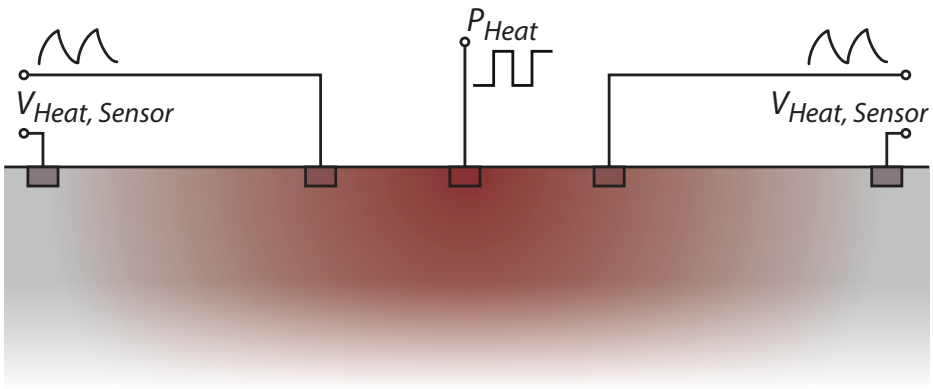

Figure 1.4: Temperature sensing principle of Thermal-Diffusivity based temperature sensors.

nologies. Moreover, since electrical delays in smaller technologies become smaller, smart TD sensors should become more accurate. The technique itself benefits from modern technologies since the thermal delay is defined by the purity of silicon combined with the heater/sensor spacing, which is defined by lithography. The temperature sensor required to sense the thermal-delayed signal can be relatively simple, since it only needs to measure an accurate delay and not an accurate temperature, and good time references are usually already present in high performance digital circuitry. All of these factors make thermal diffusivity sensors a promising candidate for thermal monitoring of SoCs.

In order to readout the delayed thermal signal a temperature sensor is needed, however this does not have to end in a 'catch-22' (a vicious loop). By using a thermopile, a differential heat measurement can be performed (see figure 1.4); the 'hot' junction picks up the time delayed heat signal while the 'cold' junction, which is very far away from the heater, stays still. Therefore in TD sensors a thermopile is often used as the sensing element. An added advantage is that multiple thermopiles can be linked together to boost the output signal. The Seebeck coefficient for Silicon and Aluminum is 440 µV/K and 3.5 µV/K respectively, thus a thermopile made from these materials will have a sensitivity of 436.5 µV/K per arm. The thermal diffusivity D itself can be approximated by [8]:

$$D \propto 1/T^n \tag{1.3}$$

where $n \approx 1.8$. As explained this signal is picked up by the 'hot' junction, after which the phase is digitized and then converted to a temperature value. If the heater is driven by a constant frequency then the conversion of temperature to phase is approximated by [8]:

$$\Phi_{ETF} \quad \propto \quad s\sqrt{f/D} \quad \propto \quad s \cdot T^{n/2}\sqrt{f} \tag{1.4}$$

where s is the spacing between the heater and the 'hot' juctions, $n \approx 1.8$, T is the temperature and f is the frequency at which the heater is driven.

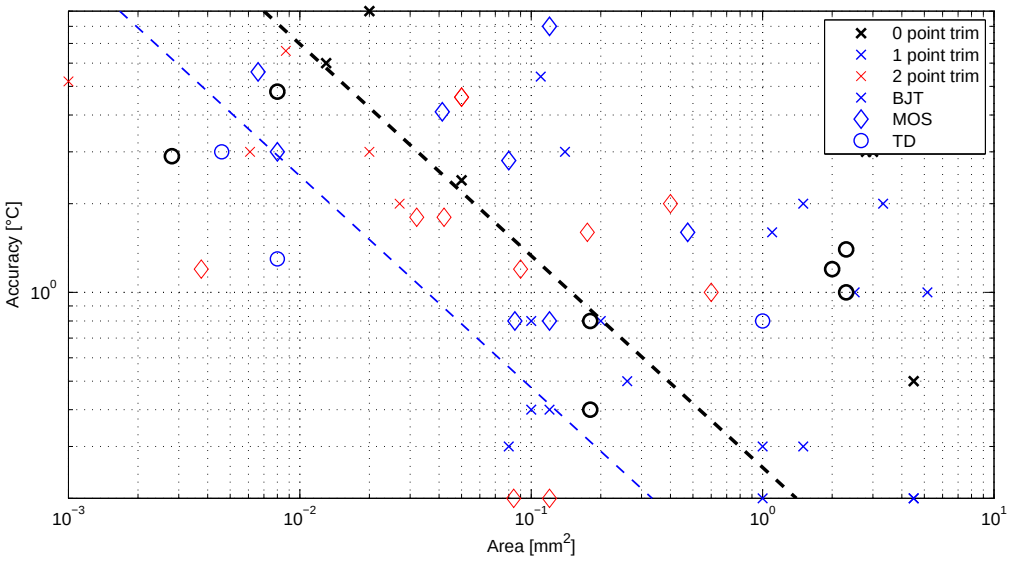

Figure 1.5: Accuracy vs. Area barrier, plotted from a survey on temperature sensors taken from [9]. Black is 0-point trim, blue is 1-point trim, red is 2-point trim (or more). The black line denotes the trend for 0-point trim BJT and MOS devices, the blue line denotes the same for 1-point trim.

1.2.6. Accuracy vs. Size

Last but not least is the accuracy-size trend. To get a certain accuracy in BJT based temperature sensors a certain area is required. This is because BJTs rely on both lithography and doping for their electrical characteristics. While lithography improves very well for newer technologies doping improves at a much slower rate, facilitating the trend of accuracy vs. area. This can be seen in figure 1.5, where a list of temperature sensors from [9] has been plotted. For zero or single point trim only TD sensors punch through the accuracy-area trend.

1.3. Outline

TD temperature sensors targeting thermal management have been made before, most notably [10] and [11]. While [10] uses an analog Gm-C integrator, which requires a large integration capacitance and analog techniques such as gain-boosting to achieve output impedance requirements, [11] is a digital reincarnation, employing a VCO counter combination to replace the analog Gm-C integrator. This approach fits the CMOS scaling story better because it removes the need for the highly analog Gm stage as well as the large integration capacitance, thus reducing the required area.

However in the transition from analog Gm-C integrator to digital VCO counter integrator, some of the performance was lost, see table 1.1. As can be seen in the table there is a big area gain, however, the accuracy and noise take a significant hit. Also, the circuit power consumption went up due to the digital implementation. Considering that for maximum SNR, most of the sensor's power should be used to generate heat pulses, the need to reduce circuit power becomes very clear.

The goal of this research project is to find out where the loss in performance comes from, and then to restore the performance of the analog approach. There is also an additional requirement not to increase the sensor area compared to [11], and if possible to reduce it.

Table 1.1: TD Sensor evolution from analog readout to digital readout.

	[11]	[10]
Area [mm²]	0.0046	0.008
Inaccuracy, No Trim [°C]	6.5 (3σ)	2.4 (3σ)
1-pt. Trim [°C]	1.5 (3σ)	0.65 (3σ)
Temp. Range [°C]	-10 — 125	-40 — 125
Resolution [°C]	0.6	0.21
Speed [kS/s]	0.9	1
Supply Sensitivity [°C/V]	-	1.3
Power [mW]	3.6	3.1

The outline of this thesis is as follows; In chapter 2 the readout, and hence the VCO-based integrator, will be studied in more detail. Then in chapter 3 the design, and hence improvements over previous TD temperature sensors, will be explained. The resulting measurements and explanations thereon will be given in chapter 4. Finally in chapter 5 a conclusion will be drawn and advice for future work will be given.

2

Phase-Domain ΣΔ-Modulator

As discussed in the previous chapter, the readout of a TD sensor involves a time measurement, or more specifically, a phase measurement. Depending on the temperature, the thermal diffusivity of silicon will change, and so a periodic heater signal will result in an output signal with changing phase. In order to determine the temperature, this phase needs to be digitized, for this a phase-domain sigma-delta modulator (PDΣΔM) is used. This chapter gives a brief description of the PDΣΔM (for more information on the PDΣΔM the reader is referred to [12]) with the aim of analyzing quantization artifacts from the VCO-based PDΣΔM. First a brief description of the operation of a TD sensor will be given.

2.1. ETF

To measure the temperature using thermal diffusivity, first a heat signal needs to be generated in the silicon, then the delayed temperature signal needs to be picked up. The sensor which does this is called an ETF (Electro Thermal Filter). It typically uses a small diffusion resistance as the heater. Then a number of thermocouples are laid out in a circle around the sensor. Each thermocouple consists of a hot and a cold junction with a non-silicided diffusion resistor in between. The junctions are the contacts between the silicon diffusion resistor and the aluminum metal. If the cold junctions are at a sufficient distance they will not pick up any heat signal from the local heater, then all the heat signal generated will be fully picked up by the thermocouples. Finally, a phase readout can be done to determine the thermal diffusivity, and with that the temperature of the die can be determined.

2.2. Ideal Phase-Domain ΣΔ-Modulator

The most interesting part of the information in the ETF output signal is in its phase. Because of the low ETF output voltages corresponding to the burned power in the heater, and because of the thermal noise of the thermopile resistance, this signal is very noisy. Therefore using edge detection for the phase readout is generally not a good idea [13]. It is better to use a measurement system in which the entire output signal of the ETF is used. This can be done by synchronously demodulating the ETF signal by a periodic signal with an adjustable and known

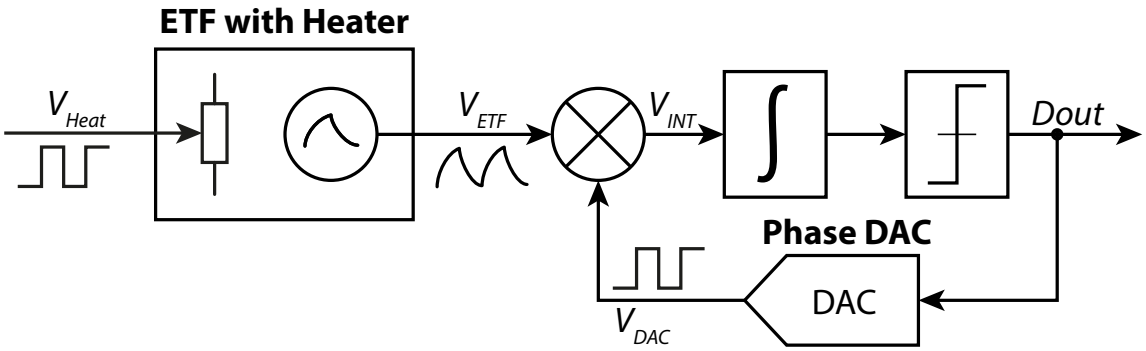

Figure 2.1: Ideal Phase-Domain Sigma-Delta-Modulator Readout.

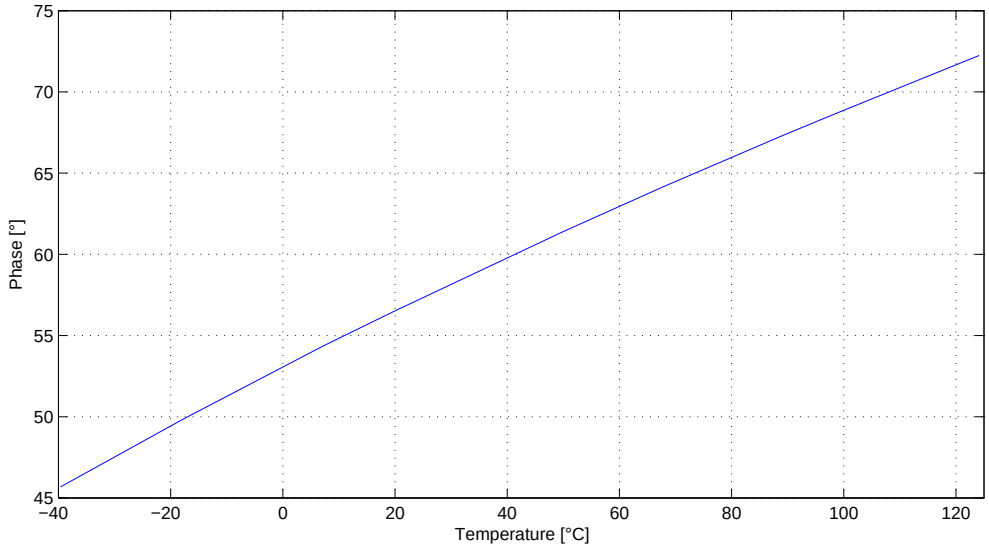

Figure 2.2: Typical ETF phase response.

phase. For two sinusoidal waveforms at the same frequency we get:

$$\cos(2\pi f t + \phi_{ETF}) \cdot \cos(2\pi f t + \phi_{demod}) = \frac{1}{2}\left[\cos(\phi_{ETF} - \phi_{demod}) + \cos(2 \cdot 2\pi f t + \phi_{ETF} + \phi_{demod})\right] \tag{2.1}$$

Thus there is a tone at twice the frequency and a DC term. If we integrate this signal for n periods the tone at twice the frequency will be filtered out; we are only left with the DC term. The DC term is dependent on the phase difference between the ETF phase and demodulation phase by a non-linear cosine function. Normally when having non-linear functions the effects of the non-linearity can be mitigated by using an actuator in a feedback loop which drives the input to the non-linear part back to a 'zero' value. In this case having a phase-DAC in a closed ΣΔ loop has the advantage that the output is digitized at the same time (see Figure 2.1).

The ΣΔ loop drives the demodulated output V_{err} back to zero. Since:

$$V_{err} \propto \cos(\phi_{ETF} - \phi_{demod}) \tag{2.2}$$

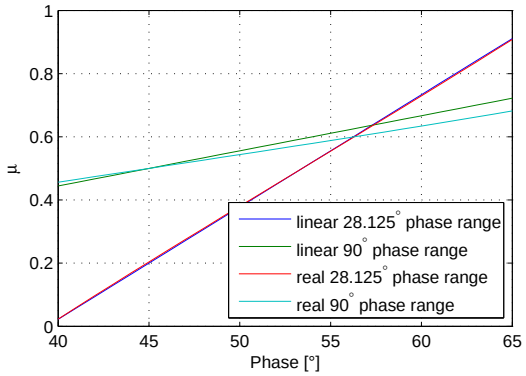

Figure 2.3: Bitstream-phase characteristic for the linear approximation and real case, where the 28.125° range has reference phases ϕ_1 = 39.375° and ϕ_2 = 67.5°, and the 90° range has reference phases of ϕ_1 = 0° and ϕ_2 = 90°.

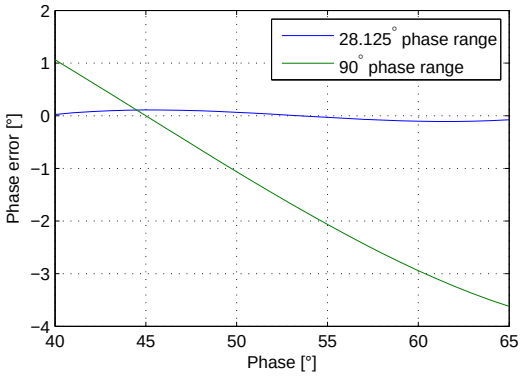

Figure 2.4: Cosine non-linearity phase error, where the 28.125° range has reference phases ϕ_1 = 39.375° and ϕ_2 = 67.5°, and the 90° range has reference phases of ϕ_1 = 0° and ϕ_2 = 90°.

this means that ϕ_{demod} will be driven to a 90° phase-shift compared to ϕ_{ETF}. However this requires the phase-DAC to have sufficient ENOB to handle the full range of ETF phases and to accommodate a small enough phase resolution to mitigate the cosine non-linearity (2.1). This means that every temperature sensor would require its own full phase-DAC, while if we were to take two phase references to straddle the temperature range then every sensor could use these two phase references. In a 160-nm CMOS technology a phase DAC requires a significant amount of area, taking into account the area used in previous TD sensors (see [10, 11]).

On second thought, however, the non-linearity associated with the use of a 1-bit phase DAC is not a big problem. When making a TD temperature sensor, a non-linear 'master-curve' needs to be generated since the ETF's temperature-to-phase characteristic is not linear anyway (1.4), see figure 2.2 for a typical ETF phase response. Since the non-linearity of the demodulation is fully deterministic it will show up in this master-curve and be mitigated when applying it. The non-linearity arises because the charge-balancing condition at the input of the integrator requires that:

$$\mu \cdot \cos(\phi_{ETF} - \phi_0) + (1-\mu) \cdot \cos(\phi_{ETF} - \phi_1) \approx 0 \tag{2.3}$$

$$\mu \approx \frac{\cos(\phi_1 - \phi_{ETF})}{\cos(\phi_{ETF} - \phi_0) + \cos(\phi_1 - \phi_{ETF})} \tag{2.4}$$

For ϕ_0 and ϕ_1 close to a 90° phase shift with ϕ_{ETF} the non-linearity can be neglected and (2.4) can be approximated as:

$$\mu \approx \frac{\phi_1 - 90° - \phi_{ETF}}{\phi_1 - \phi_0} \tag{2.5}$$

$$\phi_{ETF} \approx \mu \cdot (\phi_0 - \phi_1) + \phi_1 - 90° \tag{2.6}$$

Figure 2.3 shows how the real case (with the cosine non-linearity) looks compared to the ideal linearized case. It does this for two different phase reference settings; a full range of 90°and a smaller range of 28.125° which straddles the temperature range of the sensor. As can be seen in figure 2.4, when setting the two phase references close to the complete phase range of the

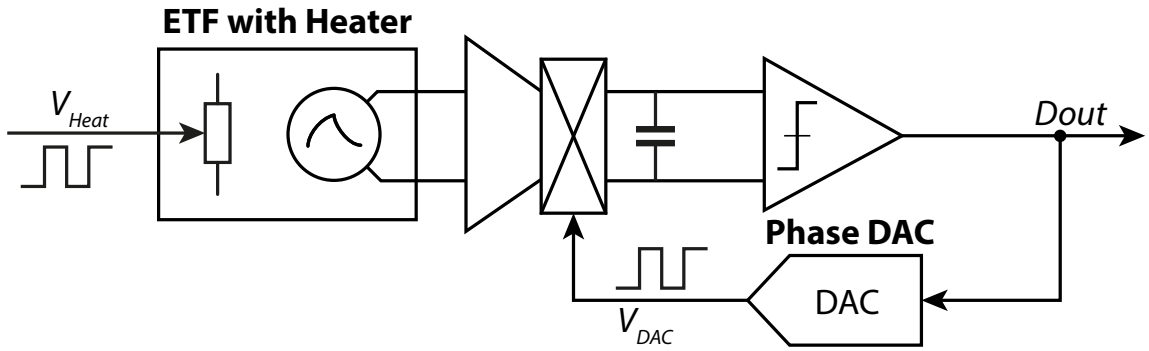

Figure 2.5: Analog Phase-Domain Sigma-Delta-Modulator Readout.

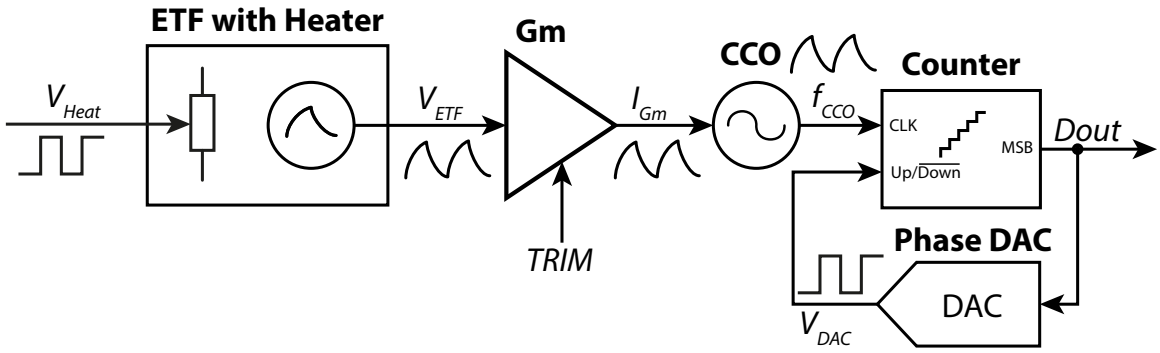

Figure 2.6: VCO-Based Phase-Domain Sigma-Delta-Modulator Readout.

sensor for the measured temperature range, the error made by the cosine non-linearity is very small.

In [10] an analog implementation of the figure 2.1 system has been made (see figure 2.5). The downside of this system is that in order to reach the output impedance requirements necessary to avoid excessive integration leakage a telescopic gain-boosted amplifier was needed. Due to the shrinking voltage headroom, this does not port very well to smaller technologies. Also the capacitor, which is about 4.5 pF, takes up a significant amount of area. That is why a more digital VCO-based front-end was designed in [11].

2.3. VCO based Phase-Domain ΣΔ-Modulator

The VCO based readout uses a VCO and counter combination (see figure 2.6) as the integrator implementation of figure 2.1. Effectively what happens is that V_{ETF} is transformed into I_{Gm} by the Gm-stage, which is then transformed into a frequency signal f_{CCO}, which is finally integrated by the counter. The demodulation is done by the up/down action of the counter. Qualitatively this makes sense because a higher V_{ETF} voltage means a higher current, and hence a higher frequency. A higher frequency on the counter clock means that the counter counts at a higher frequency and has therefore a steeper slope.

Note that this system is highly digital, and scales very well to modern technologies. In fact the only analog part remaining is the Gm-stage. In the new digital readout the counter fundamentally can't leak; which means that the output impedance requirements on the Gm

stage are much relaxed. Therefore a Gm-stage requiring less voltage headroom, which also scales better to smaller technologies, can be employed.

3

System Design

For this project two tape-outs have been done. The first one is primarily to improve the VCO-based readout. After this tape-out there was still a performance gap between the new VCO-based readout and the old analog readout. The analog readout used a larger area ETF, to find out if this was the source of the increased inaccuracy and to measure batch-to-batch spread a new tape-out was done.

3.1. ETF

As discussed in the previous, the ETF acts as both a heater and a heat detector. A periodic heating signal locally heats up the die. After a time delay, which depends on the absolute temperature of the silicon, this heat signal is picked up by the heat sensing element, i.e. the thermopile. The thermopile works based on the Seebeck effect, which states that two bi-metalic junctions at different temperatures generate a net difference in electromotive force (EMF). In our case the junction between the silicon and the aluminum interconnect can be considered as this bi-metalic junction. This approximation works since the Seebeck coefficient of metals are very similar, and differ largely from the Seebeck coefficient of silicon.

Figure 3.1 shows a cross-section of the ETF. The diffusion resistances of both the heater and the thermopiles are shown in yellow. The complete thermopile is made up of a stack of 16

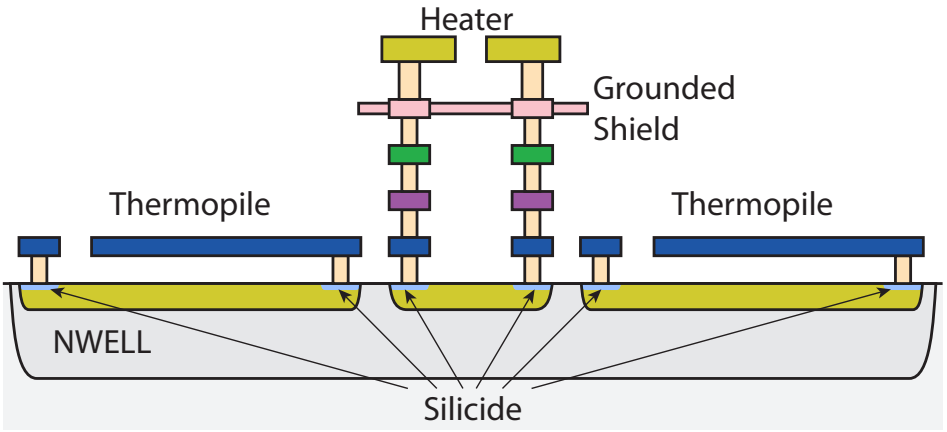

Figure 3.1: ETF cross-section.

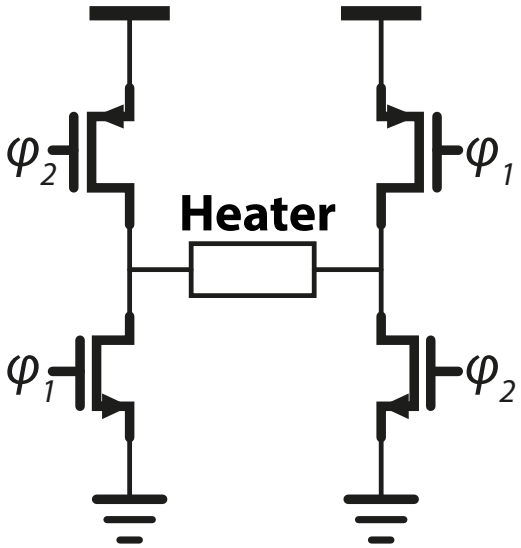

Figure 3.2: Heater resistor with switches.

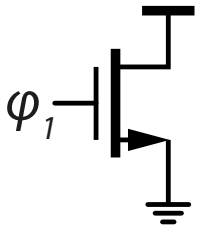

Figure 3.3: MOSFET Heater.

thermopile resistances to boost the output signal. In the cross section two of these thermopile resistances are shown. The metal interconnects must be as short and low capacitance as possible to minimize the electrical delay. Also the electrical coupling between the heater and thermopile interconnect must be as small as possible to reduce injected phase-shift from the heat drive signal. Therefore the heater interconnect is routed at the highest metal layer and a grounded shield is routed underneath, in figure 3.1 the heater interconnect is routed over the ETF (at metal layer 5) into the back of the paper, at metal layer 4 the grounded shield, which is behind and separate from the vertical columns of the heater interconnect, also runs into the back of the paper.

The ETF has 5 connections, two for the heater power, and three for the thermopile. The two outputs of the thermopile are directly connected to a differential pair, and to satisfy its common-mode requirements the middle of the thermopile is biased with a common-mode signal. A top-view can be seen in figures 3.6, 3.7, 3.9, 3.10 and 3.8. Effectively, as seen from the ETF common-mode voltage, there are two arms, each existing of 8 thermopiles, and each going to one of the differential pair input transistors.

3.1.1. Heater

Traditionally the ETF heater has always been made using a diffusion resistance. Because the temperature sensor requires a periodic heat-signal, the heater must be able to turn on and off. This means that we need switches to conduct or stop current going through the resistor. As described above, a grounded shield has been used to minimize the effect of electrical coupling between the thermopile and the heater interconnect. To reduce any adverse effects from electrical coupling the heating signal is chopped such that any DC effects will cancel out. Figure 3.2 shows a schematic of the heater and switches.

Because the resistor needs to burn quite some power (from 1.5 up to 3 mW) it also needs to conduct quite some current (1.6 up to 2 mA). Therefore the switches which need to con-

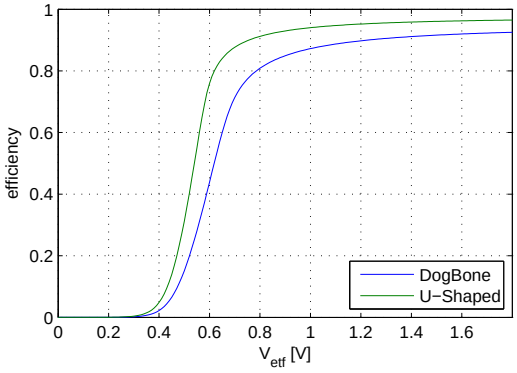

Figure 3.4: Heater efficiency.

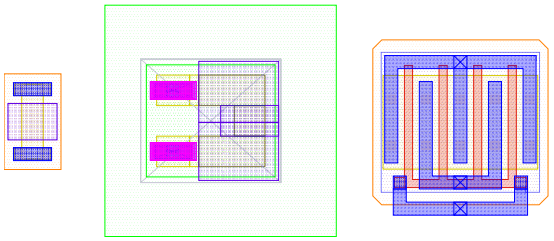

Figure 3.5: Layout comparison of the three different heaters. From left to right; DogBone, U-Shaped and MOSFET heater.

duct the heater current must have very low on-resistance, otherwise they will also start to burn power. The immediate effect is that the efficiency of the heater is reduced (the efficiency of the combination of the two different kind of heaters with the heater switches is shown in figure 3.4), however there is a second effect; thermal cross-talk. Because the temperature sensor is so small the heater switches are also necessarily located close to the ETF, and if the switches start burning power at the exact frequency of the PDΣΔM it opens up another thermal-path into the ETF. To reduce these effects the switches need to be made large, which goes against the desire to make small temperature sensors. The heater switches size is 270 μm^2, which for the smallest presented temperature sensor in this work would be almost 10% of the total area.

Obviously the self-heating of the heater switches is a problem. However, the solution to this problem is staring us in the face. Instead of trying to work around the self-heating of the switches we could also use it to positive effect, i.e. use the switches itself as the primary heat source, see figure 3.3. By using the on-resistance of the switch as a heat source the heater becomes inherently efficient, only the interconnect resistances will still limit the efficiency. Also the area-efficiency trade-of has been removed and 270 μm^2 of area has been saved. To add to that, the driving circuit can become simpler since we don't require any chopping of the power-supply and the switch can still be driven by a digital signal.

For testing purposes versions with the new MOSFET heater and two different resistive heaters; the older U-Shaped and newer Dog-Bone heater, have been taped-out. Figure 3.5 shows a layout comparison of these three heater types.

3.1.2. Square ETF

In the prior art (see figure 3.6), the hot-junctions are all at the same distance from the heater and also the cold-junctions are all at the same distance from the heater. This makes sense for the hot-junction as there is a particular distance which optimizes the trade-of for picking up enough heating power and having small phase inaccuracy. The general idea of the cold-junction is that it is so far away that it does not pickup any heat from the local heater, this however is not the ideal spot for the cold-junction. The ideal place is where it still picks-up some of the heating signal from the local heater. The effective heating signal sensed is the difference

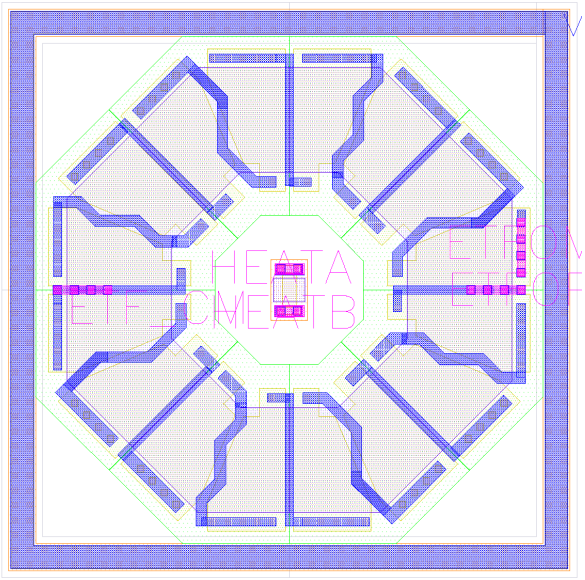

Figure 3.6: ETF7 with Dog-Bone heater.

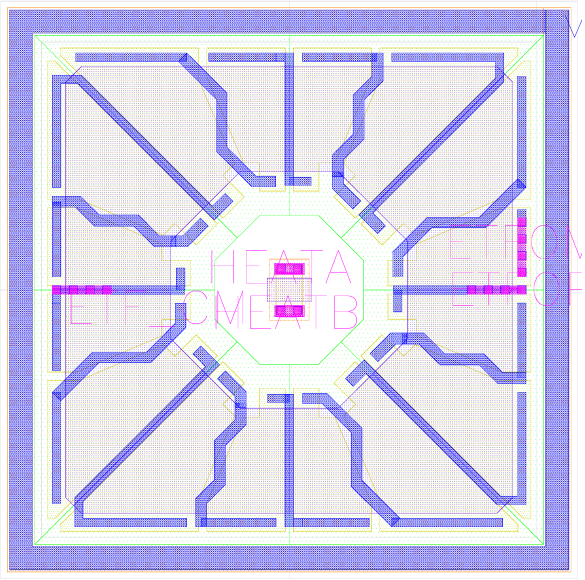

Figure 3.7: ETF7 square with Dog-Bone heater.

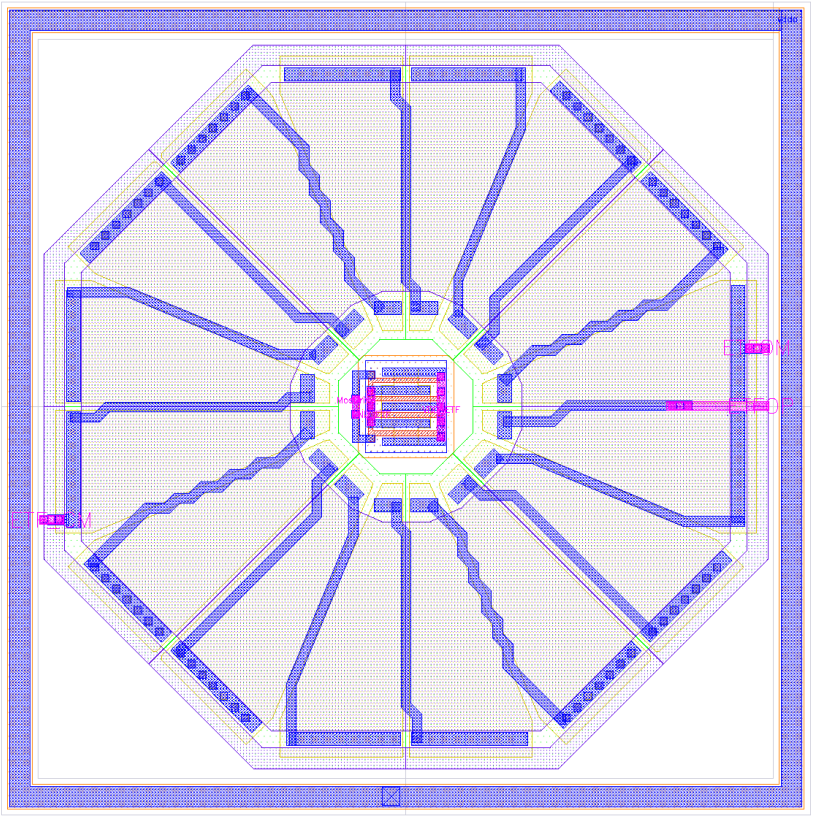

Figure 3.8: ETF3 with NMOS heater and large hot-junction.

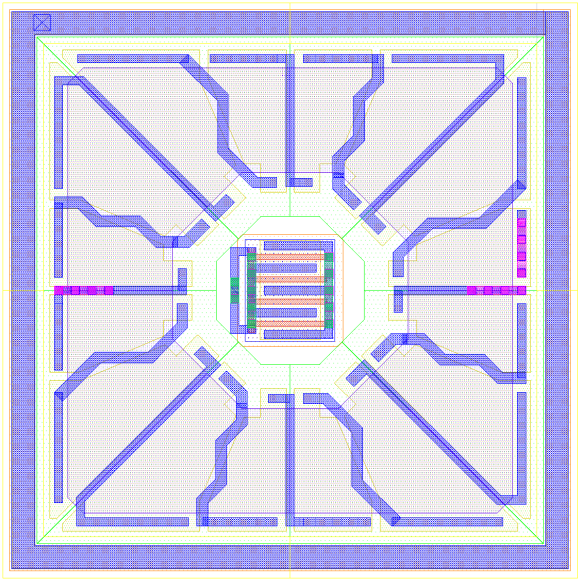

Figure 3.9: ETF7 square with NMOS heater.

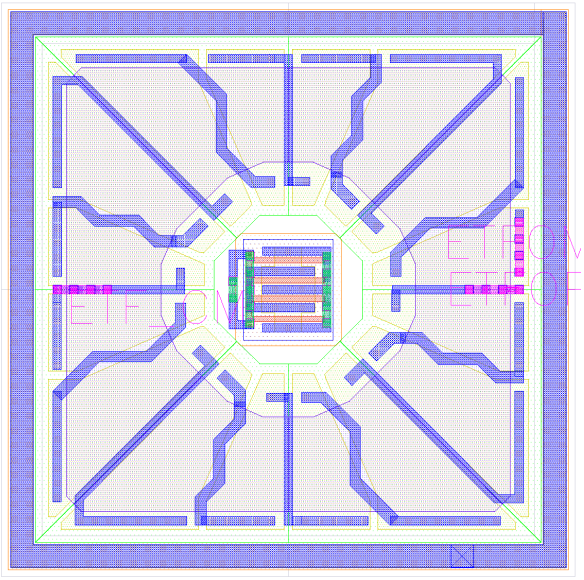

Figure 3.10: ETF7 square with NMOS heater and large hot-junction.

between the heat signal of the hot- and cold-junction. Therefore the optimum point for the cold-junction is where, if you were to extend the cold-junction, the increased noise from the increased thermopile resistance would outweigh the increase in effective heat signal. One has off-course to be careful since the cold-junction has a different phase than the hot-junction and the effective heat-signal sensed is the phasor subtraction of both the junctions. Therefore the inaccuracy in the cold-junction could, after the phasor subtraction, still have some significant effect on the total inaccuracy of the sensor.

Another factor to take into account is the area of the ETF, which is what was done in the prior art (see figure 3.6). In this ETF the cold-junctions still pickup plenty of heat signal, and can easily be further extended to get more SNR. In the prior art the ETF was purposely kept small to reduce the area, however there is no need to keep the cold-junctions all at the same distance. Therefore the cold-junctions at the corner of the ETF could easily be extended into the corners of the ETF. This will increase the SNR of the ETF while not consuming more area since this area would otherwise have been wasted. See figure 3.7, 3.9 and 3.10 for a layout of the square ETF.

In [10] a larger ETF has been used than in [11]. The larger ETF, called ETF3, was used in the analog readout, while the smaller ETF7 is only used in the digital readout. ETF7 was used as part of this work, and although the resulting sensors performed much better than [11] the inaccuracy was still larger than [10]. To rule-out the possibility that the spacing of the cold-junctions could be the cause of this increased spread, a new chip, which includes the larger ETF3 employed in the digital readout, has been taped-out. This new chip also includes versions where the silicide area is increased to see the effect of larger hot-junctions, which will be discussed in the following.

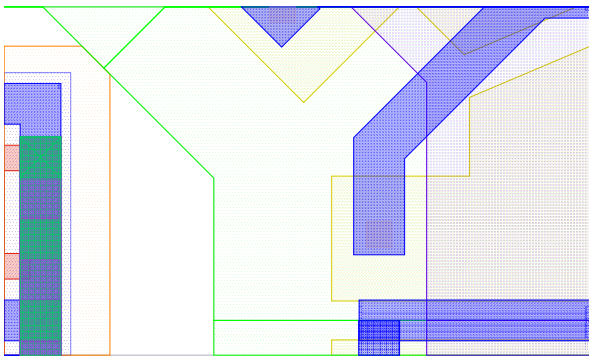

Figure 3.11: Zoom-in of the small silicide hot-junction.

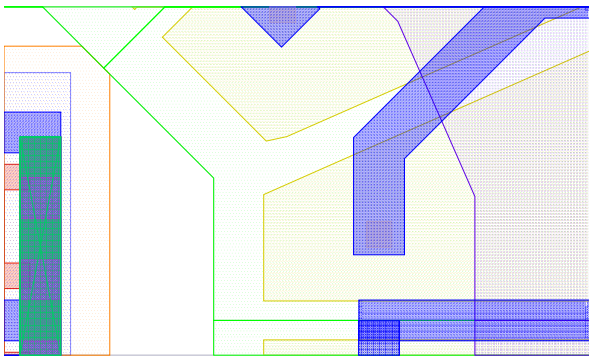

Figure 3.12: Zoom-in of the large silicide hot-junction.

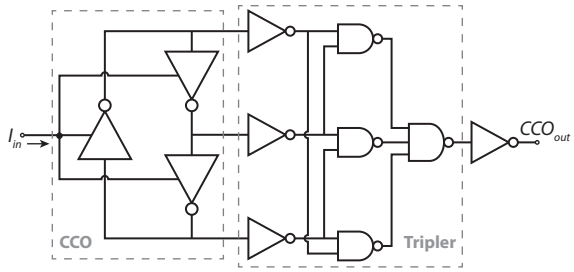

Figure 3.13: CCO plus tripler.

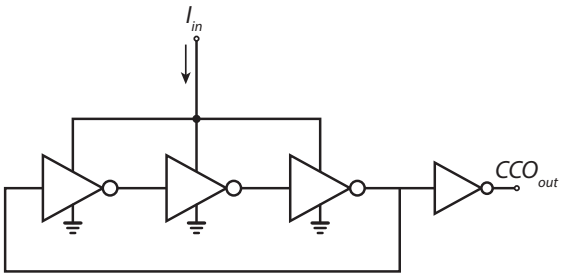

Figure 3.14: CCO plus level-shifter.

3.1.3. Silicide

The junction of the thermopiles involves metal, silicide and doped silicon (see figure 3.1). The silicide is sputtered on the silicon so that the metallic contacts don't form parasitic Schottky diodes at the contact-silicon interface. To have a large net EMF, i.e. to have an effective junction between a metal and silicon, and not an effective junction between a metal and a metal, a silicide protect layer needs to be placed over the ETF thermopile diffusion resistance.

The simple rule-of-thumb states that larger structures produce less mismatch and show less variability, because these small variations will be averaged out. To test if this is also true for the silicide junctions of the ETF we have made versions of the ETF where the hot-junction silicided area has been enlarged. This is done in two ways, first of all if you look at an ETF from prior art (see figure 3.11 for a zoom-in of the small hot-junction), you can see that the active region (in yellow) at the hot-junction has a gap. This gap is left to satisfy the DRC, because it requires the Silicide Protect layer to be orthogonal to the active where the two layers cross. By filling up these gaps and then changing the shape of the silicide protect layer we have managed to satisfy the DRC while increasing the effective hot-junction area (see figure 3.12 for a zoom-in of the new hot-junction). The second way of increasing the hot-junction area is by extending the active further in direction of the heater and extending the start of the silicide protect layer further from the heater, such that the geometrical middle of the hot-junction is still 3.3 µm away from the ETF center. For a layout of the new enlarged hot-junction see figures 3.10 and 3.8.

3.2. CCO AND LEVEL-SHIFTERS

The previous design employed a 200 MHz CCO, and a tripler to triple the output frequency, and hence tripple K_{CCO}, the CCO consisted of a 3-stage ring-oscillator (see figure 3.13). The main problem of the previous design were the level-shifters in combination with the tripler, which did not work. If the level-shifters were more carefully designed the tripler might have worked, however, at these frequencies the duty-cycle at the output of the level-shifter is very important when tripling the frequency through the use of the three digital phases of the ring-oscillator. Not even the CCO of this work, which has been very carefully designed and optimized for corner and Monte-Carlo simulations, satisfies the duty-cycle requirements for a tripler, as can be seen later.

The current to frequency gain of the oscillator is very important. In [11] the total frequency gain was severely limited because the tripler was unreliable and so had to be bypassed. This reduced frequency swing caused a lot more quantization, or discretization noise. This, in turn, was the limiting factor on the temperature resolution. Clearly it is desirable to increase the total ETF voltage to frequency gain, to reach the 90 MHz/mV for which [11] was designed.

A simple charge to frequency model of the oscillator can be made by assuming a constant swing in the CCO. Then for every period of the CCO all stages require enough charge for their gates to reach the CCO output swing. During the other half of the period the CCO stages are discharged, and don't require any charge from the input. Clearly the total capacitance of the CCO nodes is very important for the CCO frequency, and CCO frequency gain:

$$T = \frac{V_{swing} \cdot C_{tot}}{I_{in}} \tag{3.1}$$

Which is to say:

$$\Delta f = \frac{\Delta I_{in}}{V_{swing} \cdot C_{tot}} \tag{3.2}$$

To have a high K_{CCO} we want small devices with small input capacitance. The downside of this is that parasitic capacitance will be a significant factor in determining the frequency gain, and that all final optimizations must be done on layout extracted models. The upside is that since we are not using a tripler, we only need one CCO phase. This means that two of the three CCO nodes are not loaded by a level-shifter. The single level-shifter must, however, be very carefully designed; a too large load means reduced swing on that CCO node, since the unencumbered nodes run faster and don't leave enough time for the loaded CCO stage to reach full swing.

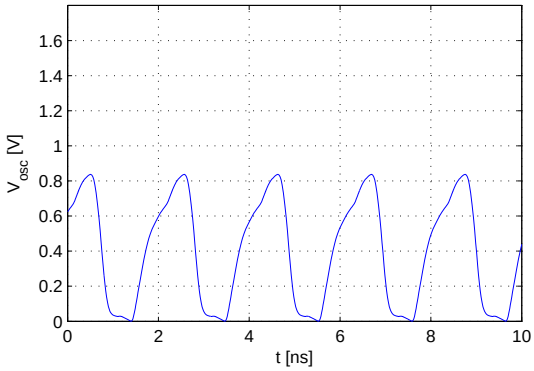

Figure 3.15: Typical output waveform of the ring-oscillator.

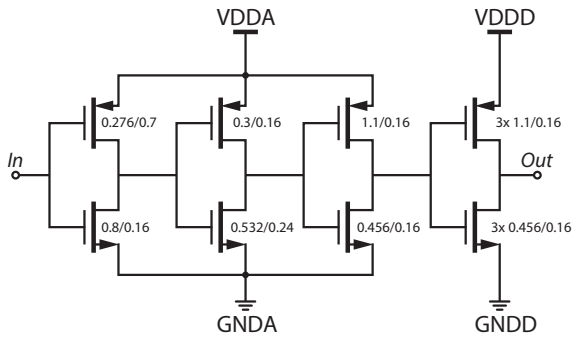

Figure 3.16: Levelshifter

Using minimum size switches for the level-shifter is not possible since the switching threshold of these devices must be somewhere in the middle of the CCO's output swing, which is below half of the 1.8 V power supply, see figure 3.15 for a typical CCO output waveform. Different techniques such as lifting the CCO's output swing away from the ground, e.g. by floating the CCO on top of a diode connected transistor, or shrinking the effective power-supply of the first stage of the level-shifter have been explored. However these techniques were either too slow (such as shrinking the effective power-supply of the level-shifter), or consumed too much area (floating the CCO on a diode connected NMOS). In the end the level-shifter as shown in figure 3.16 was chosen because it was the most robust, with smallest area, and by careful scaling of the gate-length it also achieved good input load. The first three stages of the level-shifter were connected to the analog power-supply to first regenerate the CCO signal into a strong digital signal before introducing the digital supply noise. The CCO was implemented using minimum size inverters. With the small load on only one of the oscillator stages a better current to frequency gain has been achieved, but to reach a wanted 90 MHz/mV we will need more gain from the Gm stage.

Figures 3.17, 3.18, 3.19, 3.20, 3.21 and 3.22 show the robustness of the CCO and level-shifter. These figures are the result of Monte-Carlo simulations on the Gm-stage and CCO combination after the Gm-stage and CCO combination has been trimmed to 500 MHz. Each of these figures are the result of Monte-Carlo simulations under various conditions, i.e. three different temperatures and two different input voltages which correspond to heater on and off. For all of the Monte-Carlo runs the output swing was fully from rail-to-rail, combined with the duty-cycle staying within 20% — 80% indicating that the level-shifter and CCO are working well.

3.3. Gm stage and bias circuit

Four things are important to the Gm stage, first the input referred noise must be smaller than that of the ETF. Second the 1/f corner must be below the chopping frequency of 1 MHz. Third the gain, including the CCO gain, must achieve 90 MHz/mV. And last, the inaccuracy in the added phase shift must be small.

The Gm stage in the prior art achieved all these requirements, however the PSRR was very bad, which meant that the analog power-supply needed to be finely tuned to make all tem-

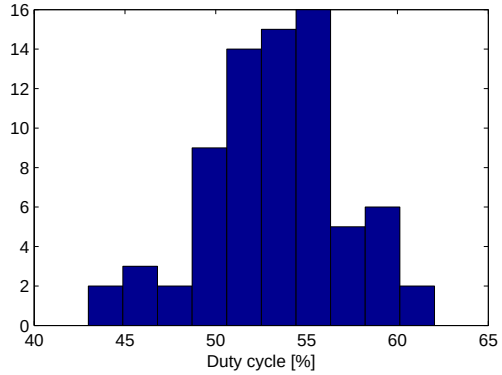

Figure 3.17: Duty cycle histogram at 125°C, low input voltage.

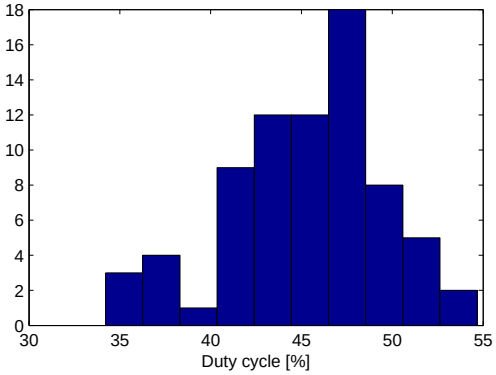

Figure 3.18: Duty cycle at 125°C, high input voltage.

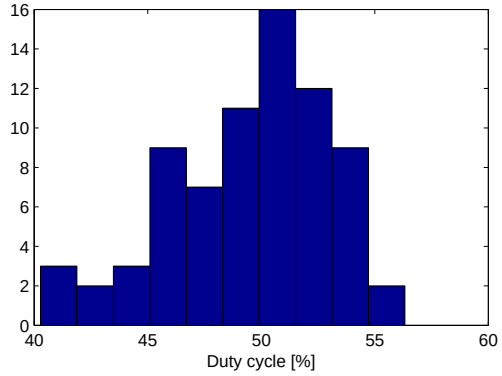

Figure 3.19: Duty cycle histogram at 27°C, low input voltage.

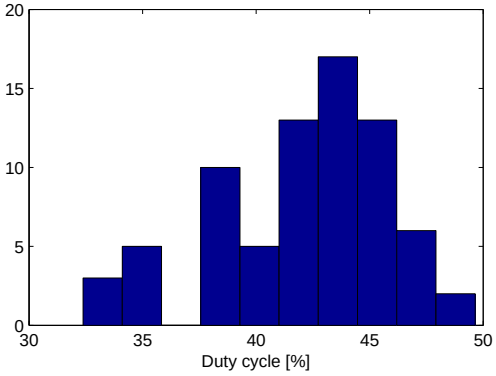

Figure 3.20: Duty cycle at 27°C, high input voltage.

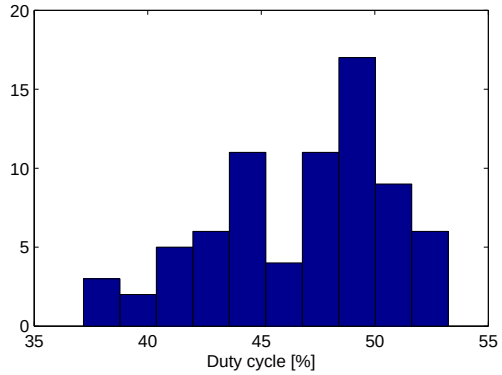

Figure 3.21: Duty cycle histogram at -55°C, low input voltage.

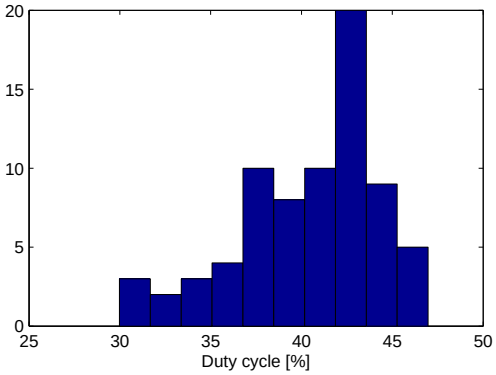

Figure 3.22: Duty cycle at -55°C, high input voltage.

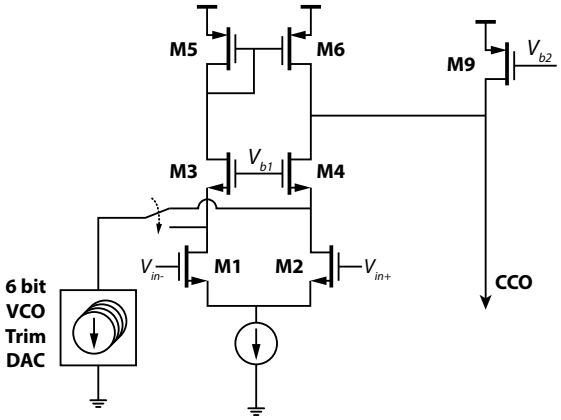

Figure 3.23: Gm-stage from the prior art.

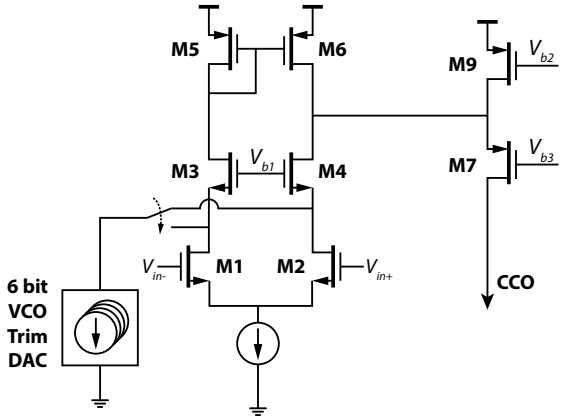

Figure 3.24: Gm-stage of current design.

perature sensors work over the temperature range. On top of that, the amplifier in prior art was designed to work in combination with the frequency tripler, thus compared to that we need more effective gain. So the previous design needed an update to achieve a VCO gain of about 90 MHz/mV, while keeping the same noise performance and phase-shift. Furthermore the PSRR needed improving to make the readout more robust and work over a larger range of temperatures.

Figure 3.23 shows the schematic of the Gm-stage of the prior art. Transistors M3 and M4 cascode input transistors M1 and M2 in order to reduce the Miller capacitance. Transistor M9 provides an offset current, so that the CCO gets an input current for zero input. A 6-bit trimming DAC is used to trim the Gm and CCO combination to a desired frequency. This is used to compensate for process variation and mismatch. Note that the CCO frequency does not need to be very accurate, hence the current DAC does not even need to be monotonic.

Figure 3.24 shows the schematic of the new Gm-stage. Compared to the previous Gm-stage it adds an extra folded-cascode. This cascode isolates the Gm-stage from the finite impedance of the CCO, thereby increasing the effective gain of the Gm-stage. This helps because the effective impedance of the CCO is reasonably large, indeed in literature the ring-oscillator is often referred to as a VCO because of this. Another small benefit is that the PSRR is improved by the cascode transistor. Transistor M7 does add another pole to the Gm-stage, and although a common-gate stage is inherently faster than a common-source stage, because the current running through transistor M7 is very small the bandwidth of transistor M7 is comparable to that of the input transistors. Simulations show that the phase-response at 1.1719 MHz was not degraded, and so to keep the circuit small and simple it was decided not to use a bleed transistor to feed transistor M7 with extra current.

The Gm-stage is biased by a constant Gm biasing circuit, which aims to keep the transistor's current gain constant over temperature. Effectively this means that the bias current depends on the temperature, see figure 3.25. The input referred noise power spectral density of the amplifier is below that of the ETF thermopile resistance, see figure 3.26. The 1/f corner frequency is just below 1.1719 MHz. Since the phase response at the ETF frequency is primarily important, the bandwidth of the amplifier must be high enough to keep readout errors due to Gm phase-response inaccuracies low. In figure 3.27 and 3.28 we can see that the bandwidth is around

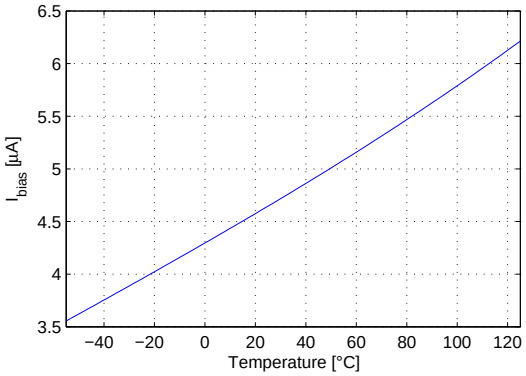

Figure 3.25: Bias current of constant-Gm-biasing over temperature.

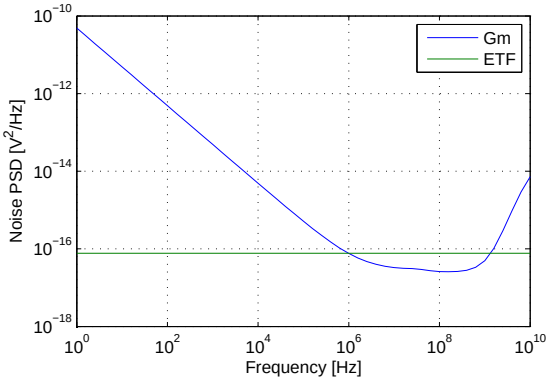

Figure 3.26: Gm-stage input referred noise.

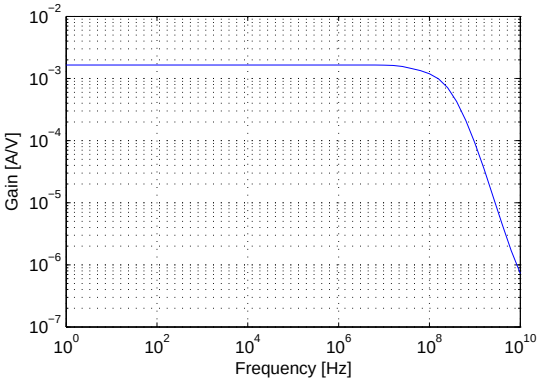

Figure 3.27: Gm-stage gain.

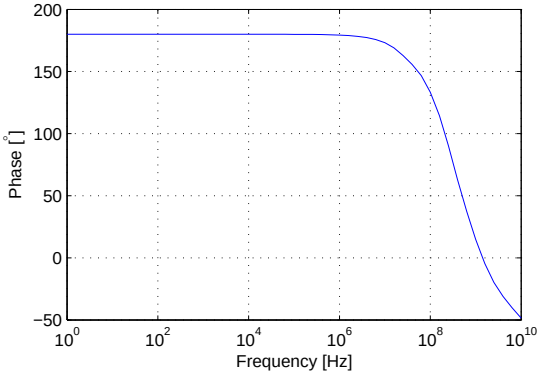

Figure 3.28: Gm-stage phase response.

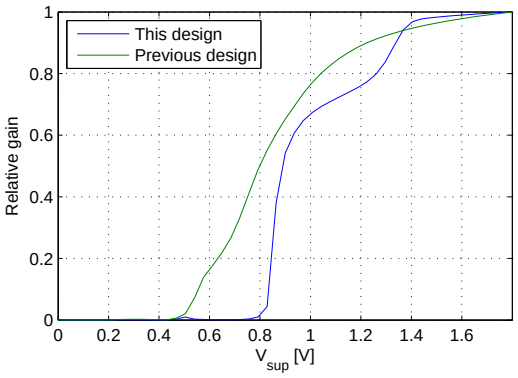

Figure 3.29: Gm gain power-supply sensitivity, at 1.1719 MHz.

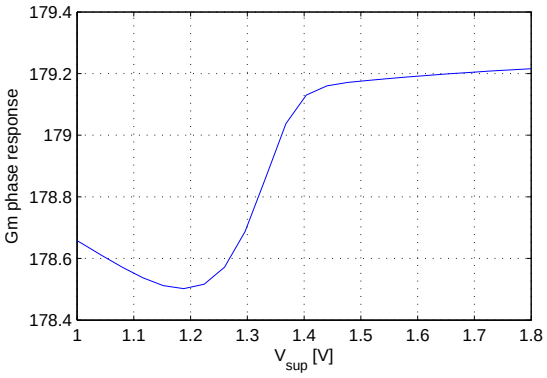

Figure 3.30: Gm phase response power-supply sensitivity, at 1.1719 MHz.

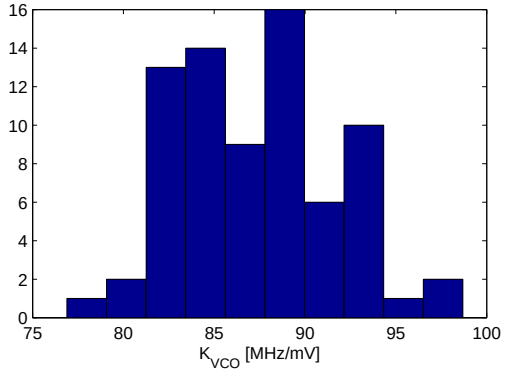

Figure 3.31: Gm + CCO gain histogram at 125°C.

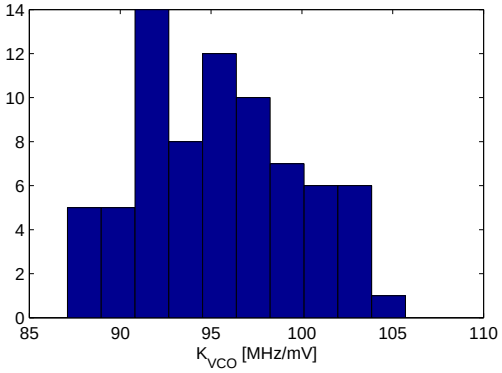

Figure 3.32: Gm + CCO gain histogram at 27°C.

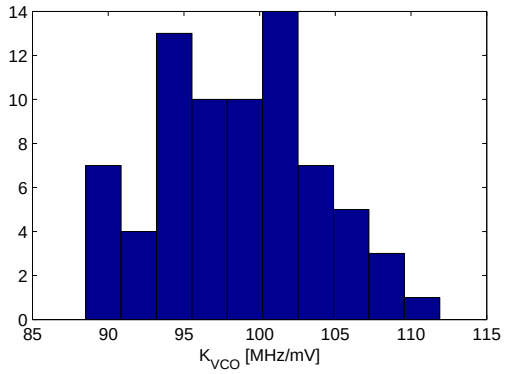

Figure 3.33: Gm + CCO gain histogram at -55°C.

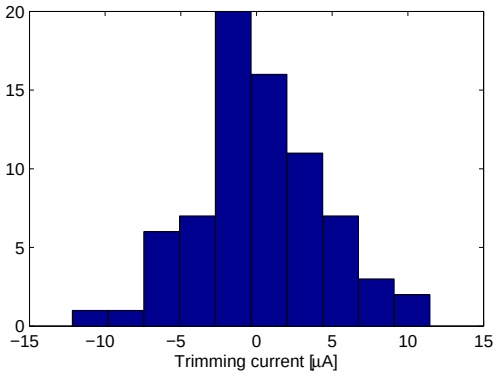

Figure 3.34: Trimming histogram at 125°C.

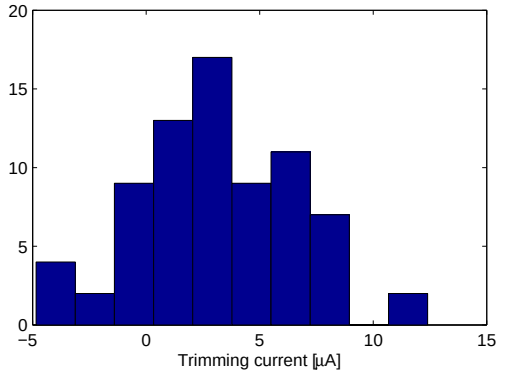

Figure 3.35: Trimming histogram at 27°C.

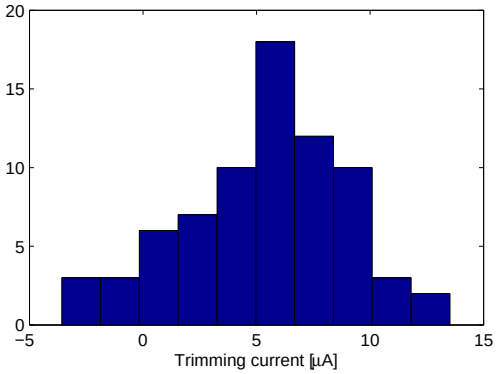

Figure 3.36: Trimming histogram at -55°C.

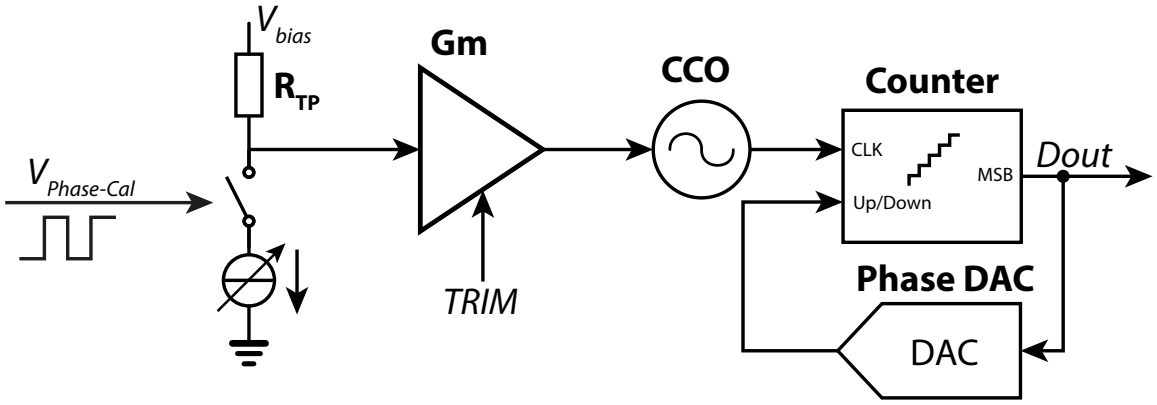

Figure 3.37: Simplified phase-calibration diagram.

100 MHz and the added phase delay at the ETF frequency is about 0.67°. From previous work, this translates into a temperature error of about 4°C, but within a batch the spread is expected to be much smaller. The PSRR has been improved by a factor 2, see figure 3.29. The phase sensitivity to the power-supply (see figure 3.30) is about 0.10°/V, which corresponds to about 0.66°C/V. The gain combination of the amplifier and CCO has been increased by a factor 3, to compensate for the loss of the tripler and to achieve a VCO gain of 90 MHz/mV. All this was done while keeping the current consumption the same.

3.4. Phase-Calibration

Every delay after the thermopile but before the demodulation, which includes the Gm-stage, and CCO and level-shifter combination, will show-up in the measured phase. If we could measure this electrical delay we could compensate for it. This so called phase-calibration will drive an electrical signal, with known phase, at the input of the Gm-stage (see figure 3.37). Then by measuring the delay as we would measure the ETF delay, we can determine the electrical delay and compensate for it. If the electrical delay spreads due to process variation or due to mismatch we can do a one-time trimming. For temperature dependent delay effects or delay drifts we can compensate with a continuous phase-calibration. Such a continuous phase-calibration requires a low noise conversion not to limit the temperature resolution.

Figure 3.37 shows the basic idea behind the phase-calibration. A periodic current is driven through the thermopile resistance, which creates a periodic voltage input at the Gm-stage. The rest of the readout is similar to a normal ETF phase measurement, only now we know the phase of the phase-calibration. After we know the measured phase we can determine the only unknown quantity, i.e. the electrical phase-delay.

To keep a simple and small biasing circuit it was decided to use only one current source, and to make the phase-calibration signal switch between the left and right arm of the thermopile, see figure 3.38. This does mean that there is a common-mode shift on the thermopile and the Gm-stage, but simulations show that the phase-calibration current was small enough not to disturb the dc-biasing condition of the ETF common-mode voltage.

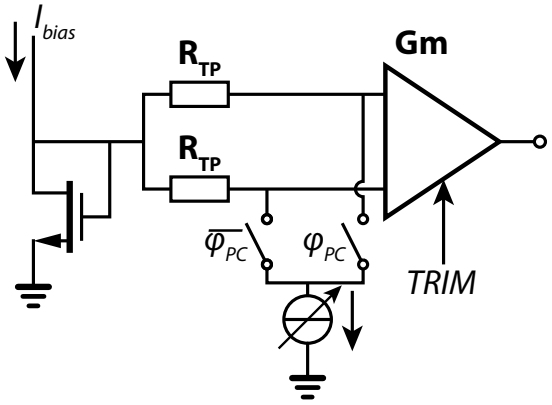

Figure 3.38: Phase-calibration schematic.

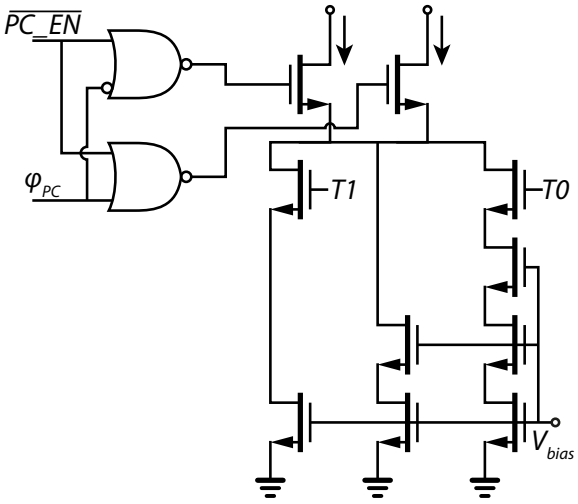

Figure 3.39: Detailed phase-calibration diagram.

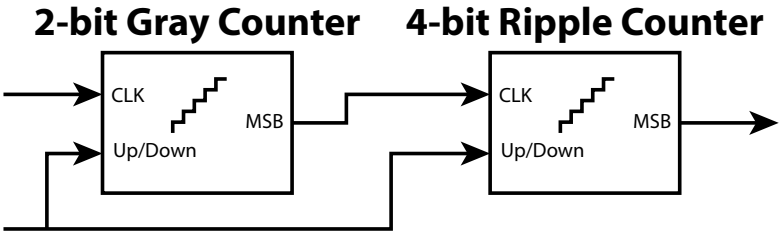

Figure 3.40: Improved low-power counter.

In figure 3.39 the detailed phase-calibration can be seen. The digital circuitry is placed locally within the analog part to reduce digital coupling into the Gm input. The phase-calibration enable and the phase-calibration phase signals come from the heater-drive. It is important to reduce any noise in the phase-calibration readout such that we can do a phase-calibration for every conversion without reducing the temperature resolution, that is why a large amplitude has been chosen for the phase-calibration signal. A current DAC has been used to make sure that a too large phase-calibration signal amplitude does not render the phase-calibration useless (see T1 and T0 in figure 3.39). The circuit has been simulated for all phase-calibration strengths, and for all strengths the amplifier is working. Simulations show that there is no increase in input referred noise or increase in 1/f noise corner due to the added phase-calibration.

3.5. Counter

In the prior art, the 6-bit counter, which ran at around 200 MHz, consumed 1.44 mW. But in this design, the timing resolution will be increased by increasing the nominal CCO frequency. This has the added benefit that the CCO works in a more linear region, and at a higher voltage swing. The downside is that, using linear extrapolation, we can expect the counter to consume somewhere around 3 mW, which would be much more than half of the total power consumption. Also, simulations show that the counter only works reliably up to 600 MHz. Therefore the counter needs to be improved to accommodate the full frequency swing under all conditions, and the power consumption needs to be lowered as much as possible.

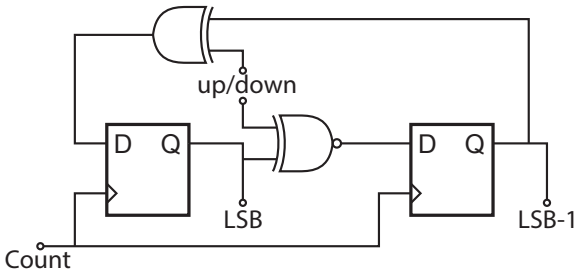

Figure 3.41: 2-Bit Gray-code up/down pre-scaler.

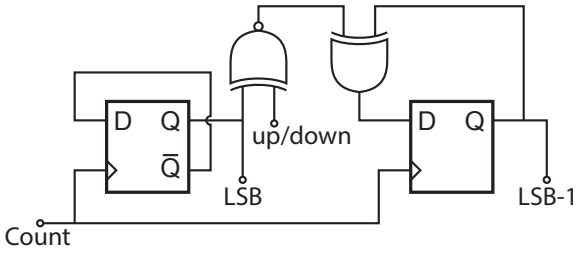

Figure 3.42: 2-Bit binary up/down counter.

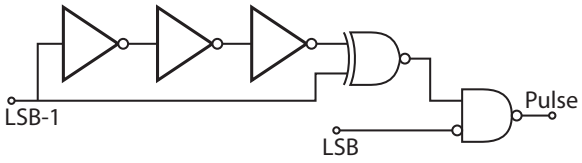

Figure 3.43: Pulse generation circuit.

To lower the power consumption the counter has been split in two parts, see figure 3.40. This saves power in the second counter because the first counter, which acts as a pre-scaler, lowers the frequency at which the second counter counts. The pre-scaler still works at the higher frequencies though, which means that this counter must be a fast counter. By making the pre-scaler a 2-bit counter we don't need to think about the carry propagation. The second counter now runs at a quarter of the CCO frequency, which means that its speed requirements are drastically decreased.

It turns out that by implementing the 2-bit pre-scaler as a Gray-code [14] counter the circuit is inherently faster and has less transitions. The reduced transitions come from the fact that Gray-code only has one bit transition per count. This means that four counts require four bit transitions, while in a binary counter four counts require six bit transitions. Thus by implementing the 2-bit pre-scaler as a Gray-code counter we have reduced the bit transitions by 30%. The speedup comes from the fact that there is no sequence of an xor and xnor gate, see figure 3.41 compared to figure 3.42.

By implementing the 4-bit counter as an binary ripple counter the power consumption can be further reduced. Every next bit in the counter will effectively run at half the frequency, which adds up to an effective clocking saving up to 50%. However implementing a 4-bit up/down ripple counter requires much more logic than for a normal 4-bit up/down counter. The problem comes from the fact that we are putting logic on a clock input, which is edge sensitive, and should be clocked when transitioning between two states. For this transition it doesn't matter if we go back or forth, in both cases the next bit should be toggled. This makes it quite hard for static logic since it should give a rising edge in both cases; going back and forth between the two states. Since this increase in logic drastically increases the counter area, and since the counter area is already quite large compared to the total sensor area it was decided to implement the 4-bit up/down ripple counter using dynamic logic, for which a pulse generator will be used.

The pulse generator will generate a pulse whenever the pre-scaler switches between the two states where the ripple counter should increment or decrement its state. The pulse generator

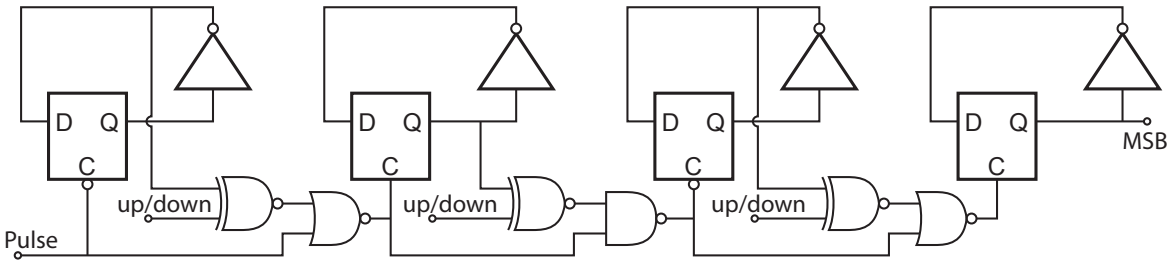

Figure 3.44: 4-Bit ripple up/down counter.

is shown in figure 3.43. It relies on an inverter delay line to generate its pulse, this is why additional care was taken during the design; Monte-Carlo simulations were done under different conditions, i.e. frequency and temperature, with extra capacitive loading on all the lines deriving from the pulse generator. It was only after increasing the extra capacitive loading to five times the parasitic capacitance that the circuit started to fail in some Monte-Carlo runs. These tests show that, even though dynamic logic was used, the counter is very reliable.

Because every register is now clocked with a short pulse, we don't need positive edge triggered flip-flops. Using level sensitive registers has been the main driving force in reducing the counter area. By careful consideration of the driving strength of all digital components and by careful consideration of which logic element combination to take, the area of the complete counter has been reduced by 20%.

Figure 3.44 show a schematic of the 4-bit ripple counter. The logic of the counter is the gating signal, i.e. the gating signal determines when the register should toggle its output. This is reflected by the fact that every register is fed the inverse of its output on its data input port. Every next stage in the counter now only needs to determine if its register should toggle, and if so let the pulse pass, otherwise it should stop the pulse from propagating. The up/down signal for the ripple counter is a buffer delayed version of the up/down signal in the pre-scaler. The up/down signal in the pre-scaler is resynchronized to the clock signal to avoid any metastability. With all the changes made, the counter works over a higher frequency range, while saving 60% of power compared to the old counter. One important thing to note is that, although not shown in figure 3.42 and 3.44, all registers have a reset port.

3.6. Heater-drive and top-level

The heater-drive controls the entire temperature sensor, not only does it generate the signal to drive the heater, but it is responsible for all other driving signals. Phase delays between the driving signals, i.e. ϕ_{ETF}, ϕ_{demod} and ϕ_{PC}, directly result in measured delays. To minimize delay differences the heater-drive has a high frequency reference clock. All important phase signals are re-clocked with this high frequency clock. For a simplified diagram of the heater-drive see figure 3.45.

The heater-drive must be able to work in three different modes; temperature conversion mode, phase-calibration mode and trimming mode. Each of these modes requires different driving signals. For the temperature conversion the ETF must be driven by a periodic signal, the phase-calibration must be disabled and the phase DAC must multiplex between ϕ_0 and ϕ_1, which

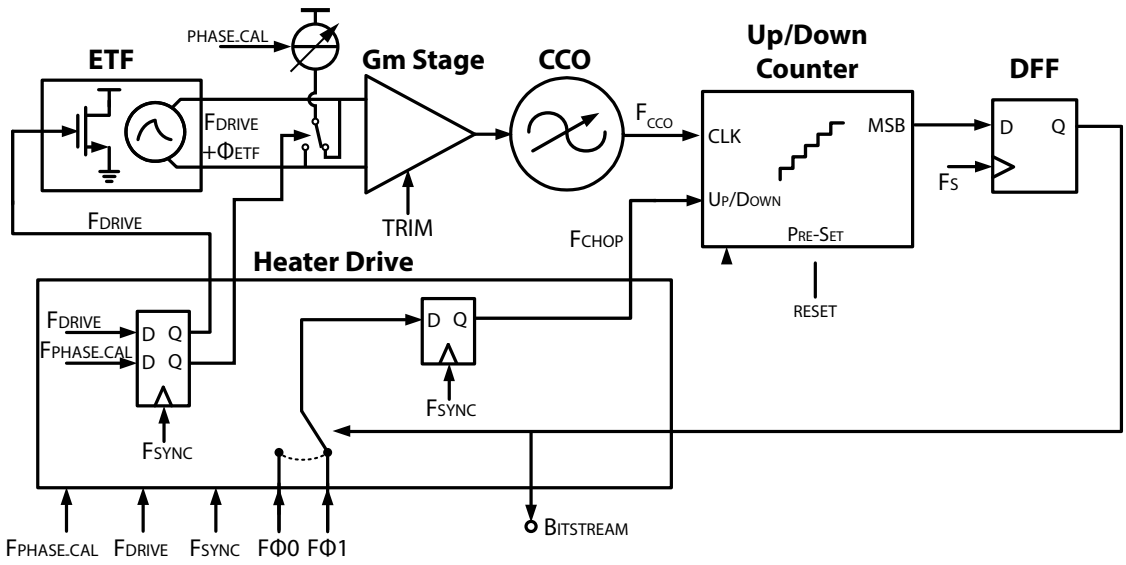

Figure 3.45: Simplified block-diagram of the temperature sensor.

are generated outside of the temperature sensor. For the phase-calibration mode the phase-calibration must be enabled and driven by a periodic waveform, while the ETF heater must be turned off and the phase DAC must multiplex between ϕ_0 and ϕ_1. For the trimming mode, to account for ETF self-heating effects, the ETF must be driven with a high frequency, such that the filtered ETF output is mostly silent, but has the DC self-heating offset. The phase-calibration must be disabled and the phase DAC must be disabled, which means that the counter only counts up. In this case the counter acts as frequency divider, the divided frequency is send off-chip where an FPGA can determine the frequency and close the trimming loop.

The bias-block and clock generation system are not include in the temperature sensor area. Because the temperature sensor is so small multiple have been placed on one die, which means that they can share these systems. This is similar to real thermal management, where multiple sensors need to monitor the multitude of potential local hot-spots. The clock generation system provides each sensor with the high-frequency reference clock, an ETF input clock and the two reference clocks for the phase DAC. For phase-calibration the ETF input clocking signal serves as phase-calibration signal and is delayed such that the phase references can be put around the electrical delayed phase-calibration signal. The extra delay on the phase-calibration is done because the electrical delay is very small and could be in the dead-zone of the PDΣΔM.

Figure 3.46 shows a detailed block-diagram of the temperature sensor (without the phase-calibration). The ETF, on the left, is connected to the Gm-stage, which in turn drives the CCO. The CCO signal is picked up and regenerated by the level-shifter and is fed to the clock input of the counter. The MSB output of the counter is sampled by the sampling clock and serves as the bit-stream. The loop is closed by selecting either of two reference phases, depending on the bit-stream value, and is used to demodulate the CCO signal, which means it is used for the up/down signal. As discussed above, all phase signals are re-clocked by a high speed reference clock, and the up/down signal is also re-clocked by the counter clock to avoid meta-stability.

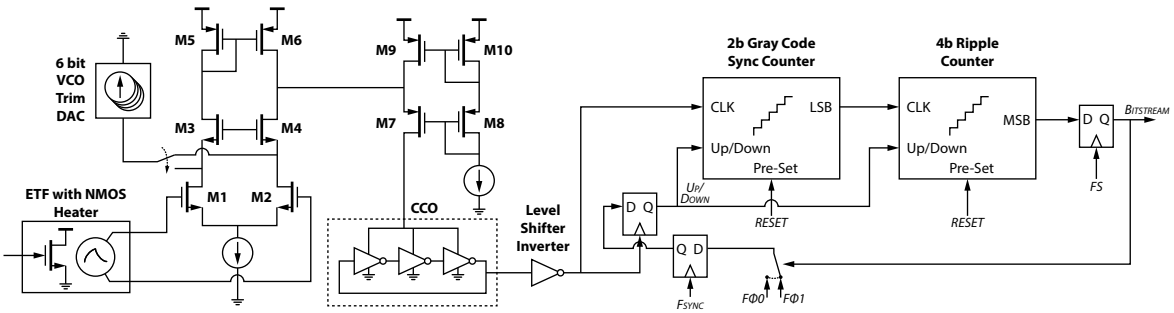

Figure 3.46: Detailed block-diagram of the temperature sensor.

Multiple different sensor readout circuits have been made, see figure 3.47. The topmost sensor is the sensor from [11]. Three different type of heaters have been made, where the difference between resistive or MOSFET heater causes different heater-drivers. The two middle sensors employ a resistive heater, while the two bottom sensors employ a MOSFET heater. The left column of sensors utilize the new low-power counter, while the right column of sensors use the old counter. Only taking square areas into account, the areas in left-to-right top-to-bottom order are: 4600 µm, 4000 µm, 4000 µm, 3600 µm and 2800 µm. Thereby an area reduction of 42% has been achieved for the smallest sensor version.

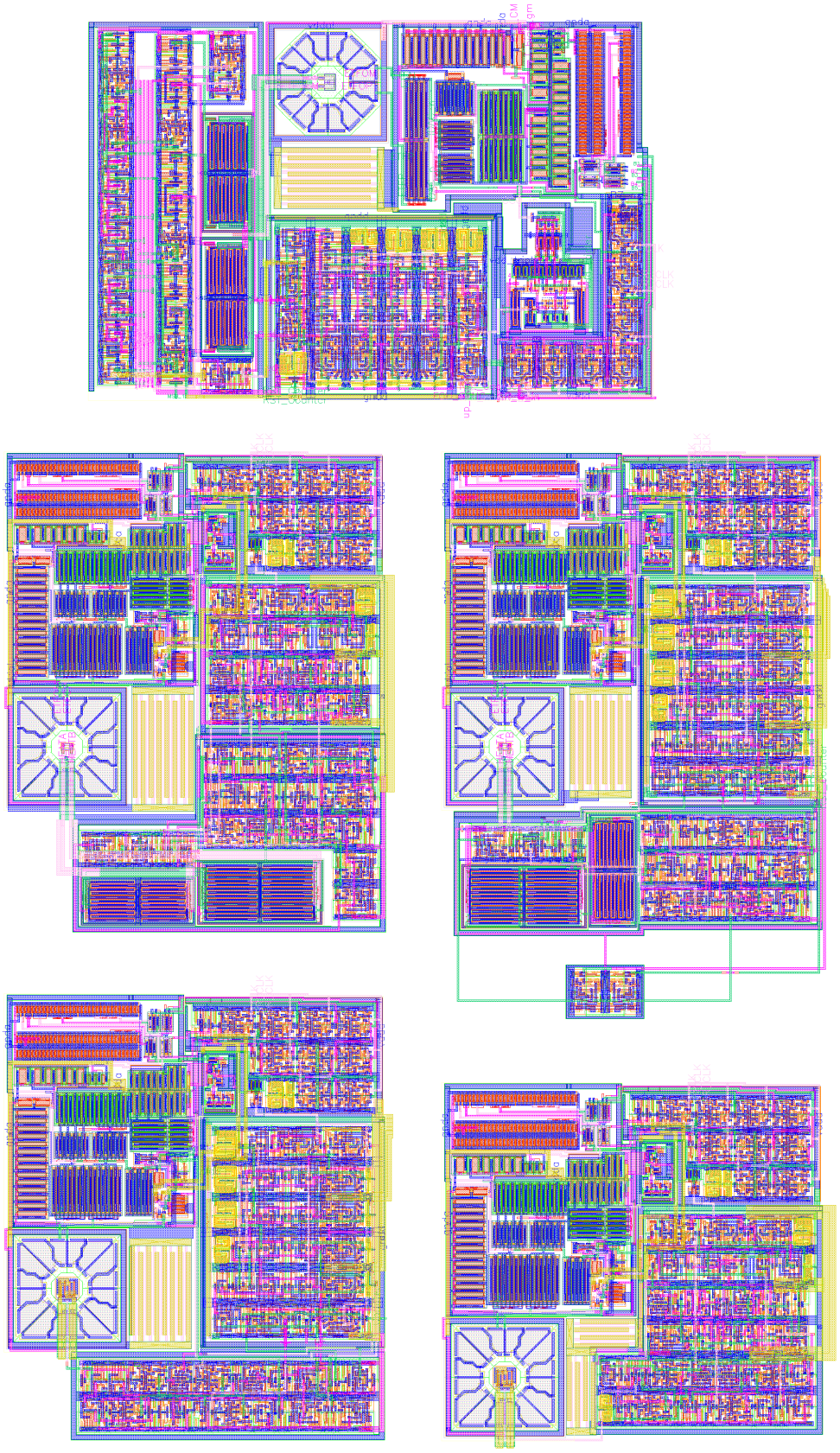

Figure 3.47: Pixel comparison of different temperature sensor designs.

4

Measurements

4.1. Measurement Setup

Since the measured chips are temperature sensors, they have to be tested over a wide temperature range. For this we need an oven and an accurate temperature sensor as reference. The oven temperature itself doesn't need to be set particularly accurately, as long as the temperature at which the chips are measured is accurately known. The temperature of the oven needs to be stable however, and to help relax this requirement a big thermal mass in the form of an aluminum block has been used. The chips are pressed against a thermal-rubber stuck to the block to improve the thermal connection. A calibrated PT-100 temperature sensor is inserted, with thermal paste applied, into a hole in the aluminum block (see figure 4.2).

To measure over a given temperature range, fixed temperature points spanning that range are chosen, e.g. 125, 105, 85, 65, 45, 25, 5, -15 and -35. During such an 'oven-run' the oven is first settled to the first temperature point, where a measurement is done, then it does the same for the second, third, ... etc. To keep errors due to the measurement itself below errors from the chips, the oven is considered settled only when the aluminum block has a delta of less than 20 mK for the duration of a single measurement (around 80 seconds).

To acquire the data, a DAQ-card (the NI-6537B) and a PC are used. This DAQ-card is then connected to a FPGA. The FPGA controls not 1, but 4 of the measured chips, to reduce the amount of oven-runs we have to do (see figure 4.1). The FPGA interfaces with the PC through the DAQ-card, the PC now only has to issue measurements through the DAQ-card and FPGA,

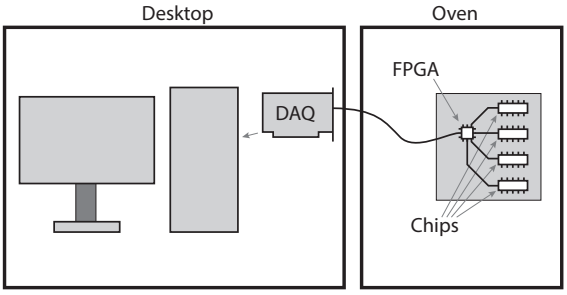

Figure 4.1: Simplified schematic overview of the measurement setup.

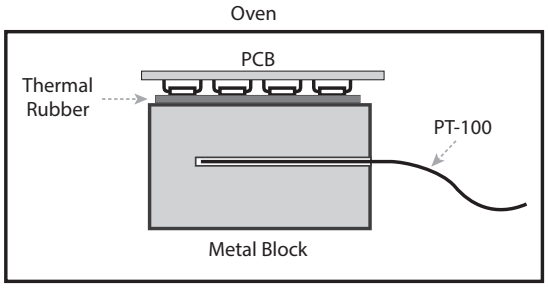

Figure 4.2: Simplified schematic of the thermal connections of the measurement setup.

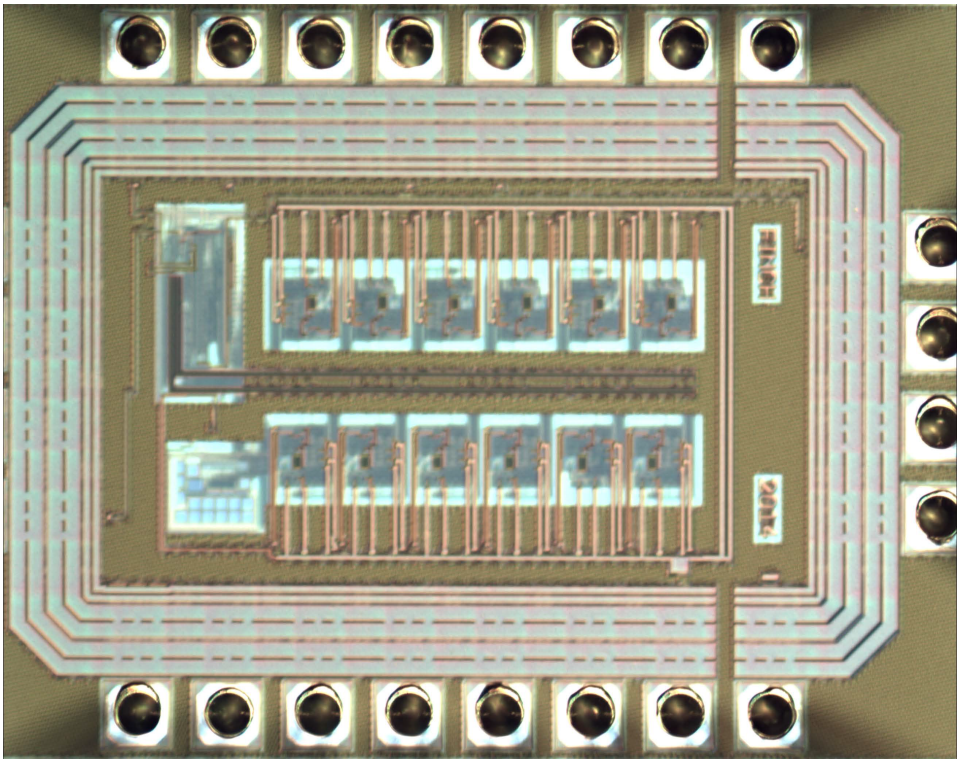

Figure 4.3: Die-Graph of the FINCH tape-out.

and wait for the results. While the FPGA is very quick and can react with clock level speeds (75 MHz), the DAQ has a lot of software and data-bus overhead. With careful programming, the delays are hidden in parallelism and a speed of 0.5 kSa/s has been achieved, which is close to the theoretical maximum of 1 kSa/s. This high speed conversion reduces the measurement time and relaxes the requirements on the thermal settling of the metal block.

For every temperature point, all relevant information is extracted to prevent unnecessary reruns. All the data consists of: a normal conversion of 1024 bits, a two-step conversion of 1024 bits where the fine references are calculated from an accompanying 64 bit normal conversion, a trim measurement of the CCO frequency at all 127 possible Trim-DAC settings, and a phase-calibration measurement. All of this is done for every pixel of all four chips.

A normal conversion consists of first a frequency trim and after that a normal 1024 bit conversion with fixed phase references. In this conversion both the bit-stream and the readout phase are extracted and saved. A two-step conversion also begins with a trim, then a 64 bit coarse conversion whose result is used to calculate and set the phase references, after which a 1024 bit fine conversion is done. In this conversion the bit-stream, phase and coarse phase are saved. A trim measurement consists of manually setting the Trim-DAC value and then counting the frequency, this count result is finally saved with the according Trim-DAC setting. A phase-calibration measurement begins with a trim, and then, using fixed phase references and for all four phase-calibration settings, a normal 1024 bit conversion. For this measurement both the phase and bit-stream for all four strengths are saved. To filter out noise/resolution errors, the phase-calibration, fine and coarse measurements are averaged 127 times.

Finally after all oven-runs are done (20 chips for both the chips with resistive heater and the

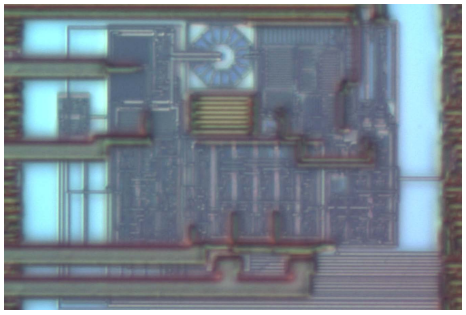

Figure 4.4: Graph of pixel with conventional counter and ETF6.

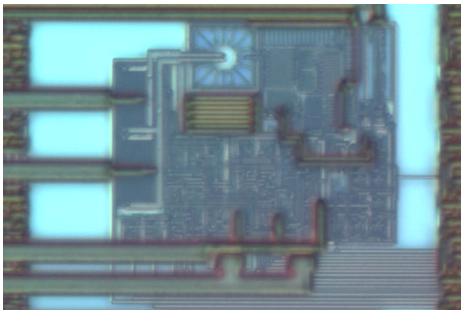

Figure 4.5: Graph of pixel with low-power counter and ETF7.

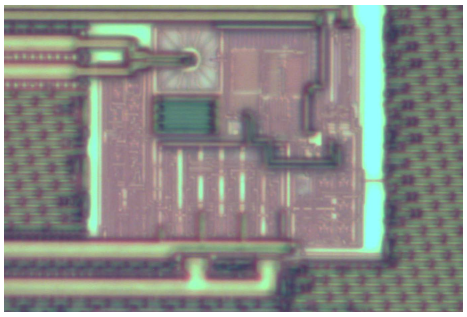

Figure 4.6: Graph of pixel with conventional counter and ETF7 with MOS heater.

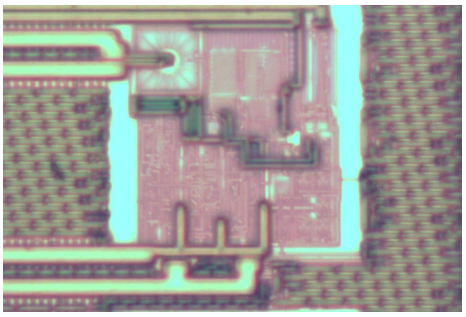

Figure 4.7: Graph of pixel with low-power counter and ETF with MOS heater.

chips with MOSFET heaters, and 16 chips for the ETF test chip) all the data was processed. The next section will explain how the post-processing was done, so that the next section (section 4.3) can simply show the results.

4.2. Master Curve

The last ingredient needed before we start with the measurements and results is the master curve. The result from the PDΣΔM is a phase and not a temperature, this is where the master-curve comes in. In short the master curve transfers a phase readout into a temperature readout. The master curve is important for most measurements done on the temperature sensor, for example when we want to know the resolution of the PDΣΔM in temperature we need first to transform the phases into a temperature, after which we can calculate the sigma in temperature (which for the noise limited PDΣΔM is its resolution). The master curve captures the behavior of the entire temperature sensor, including delays before the up/down demodulation. The most notable of all these delays is the Thermal-Diffusivity delay, which is the temperature sensitive delay we are trying to measure.

The master curve is made by doing a least-square-error polynomial fitting of all phase-temperature points, where the temperature has been taken from the PT-100 measurement. The master curve then represents a function which transforms a phase from the sensor into a corresponding temperature (4.1).

$$T_{etf} = f\left(\Phi_{etf}\right) = a_0 + a_1 \cdot \Phi_{etf} + a_2 \cdot \Phi_{etf}^2 + ... + a_n \cdot \Phi_{etf}^n \tag{4.1}$$

The order of the polynomial is important for getting correct resolution and accuracy results.

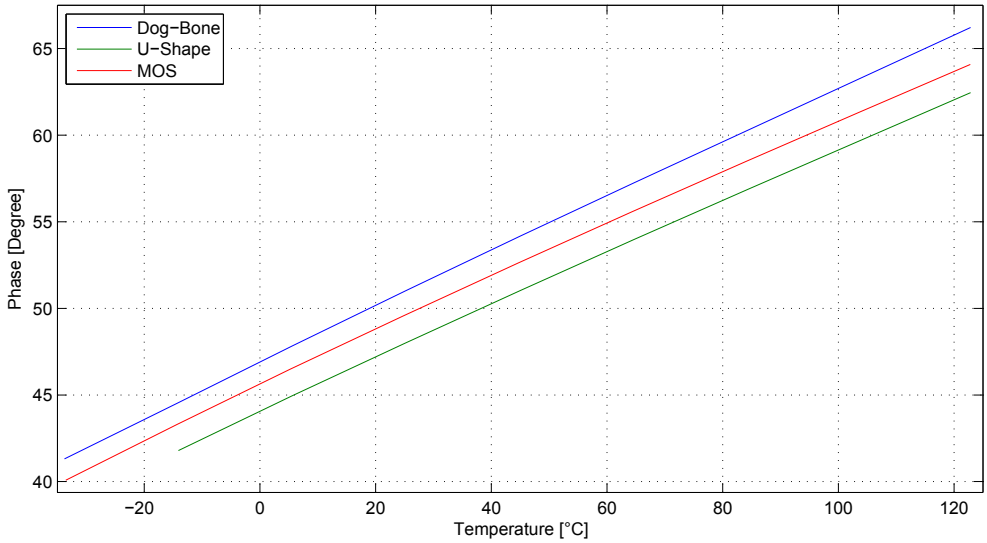

Figure 4.8: Phase read-out plots of all three ETF types.

First of all, it is important to note that the master curve is not linear, ideally the phase should actually be proportional to $T^{0.9}$. Another non-linearity factor is the cosine non-linearity, which comes from the fact that the dc term of a multiplication of two sinusoidal waves results in a cosine of the phase difference. Using the right polynomial order (5) in (4.1) is important. Taking a higher order will cause over estimation at the edges of the temperature range which causes severe distortion of the slope, or sensitivity, at those points. Taking a lower order will result in larger mean deviations of the temperature points.

Because all temperature sensors have been measured over different 'oven-runs', they will have been measured at slightly different temperatures. To be able to compare the measured phases of all temperature sensors, the temperatures at which the measurements are done first need to be homogenized. To do this the master-curve is used, this time though it is used on each individual temperature sensor. This individual master curve is then used to estimate the sensor's phase at a fixed set of temperatures.

4.3. Results

The design has been taped-out in NXPs CMOS14 process. Figure 4.3 shows a picture taken of the die with the resistive heaters, while figures 4.4, 4.5, 4.6 and 4.7 show pictures taken of the different temperature sensor versions. Each individual sensor in a die is henceforth referred to as a pixel.

In figure 4.8 the characteristic master-curves of the three different kind of ETFs are plotted. The dog-bone heater, which has the same thermopile as the rest of the pixels, has by far the largest phase shift. This can be explained by the fact that the dog-bone heater has the smallest circumference of all the heaters. This effect has been calculated by modeling the heater as a finite point source (see figure 4.9), and then calculating the expected phasors of the hot and cold junctions, see table 4.1. The MOS heater's circumference is larger, while the U-Shaped heaters circumference is the largest of all. The y-axes limits of the figure 4.8 are the phase ranges from

Table 4.1: Estimated ETF phase and voltage amplitude at 27°C, for the hot and cold junctions, and the effective thermopile output.

Heater	$\lvert V_{eff}\rvert$ [mV]	$\lvert V_{hot}\rvert$ [mV]	$\lvert V_{cold}\rvert$ [mV]	Φ_{eff} [°]	Φ_{hot} [°]	Φ_{cold} [°]
Dog-Bone	1.60	1.81	0.36	35.0	44.8	95.0
MOS	1.69	1.89	0.36	32.8	42.4	93.6
U-Shaped	1.54	1.83	0.44	32.5	44.0	88.1
Dog-Bone round ETF	*1.52*	*1.81*	*0.44*	*33.2*	*44.8*	*88.8*
MOS round ETF	*1.61*	*1.89*	*0.45*	*31.1*	*42.4*	*87.4*
U-Shaped square ETF	*1.62*	*1.83*	*0.36*	*34.2*	*44.0*	*94.3*

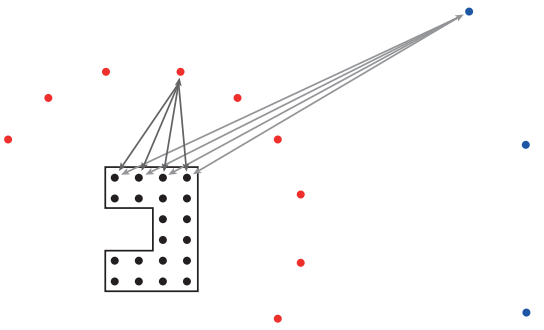

Figure 4.9: ETF characteristics approximation by modeling the heater as a finite point source.

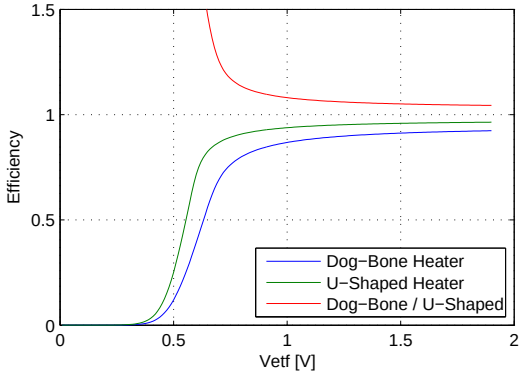

Figure 4.10: Efficiency of resistive heaters, which is power burned in the heater divided by power burned in the heater and switches. The red line is the efficiency of the U-Shaped heater divided by the efficiency of the Dog-Bone heater.

the DAC, which is 39.375° to 67.5°. The absence of data from the U-Shaped heater at -35°C is thus explained by the limited range of the phase DAC.

4.3.1. ETF Power

The MOSFET heater chips were measured first and while finding the best configuration of the chips it was decided that the ETF should be run at a similar power as previous art [11]. However, it was found that at higher heating powers the PDΣΔM loop sometimes becomes unstable. This always happens at the more extreme temperatures when one of the phase references is furthest away from the ETF phase, thereby causing a large counter swing and eventually wrap-around. Therefore it was required to either reduce the phase range – which in turn reduces the temperature range–, do a multi-step coarse conversion or reduce the power to safer levels.

In the end we have decided to reduce the power. Because of the lower heater power the resolution is dominated by the ETF noise, even for a simple 1-bit conversion. In this way the total design becomes much simpler, not only because the control logic becomes easier, but also since all pixels now require the same reference. This reduces the necessity of a phase DAC for every pixel thereby greatly reducing the complexity and surface area requirement for a complete thermal management temperature sensor.

For the resistive heaters it makes sense to either work at the same power, or at the same noise

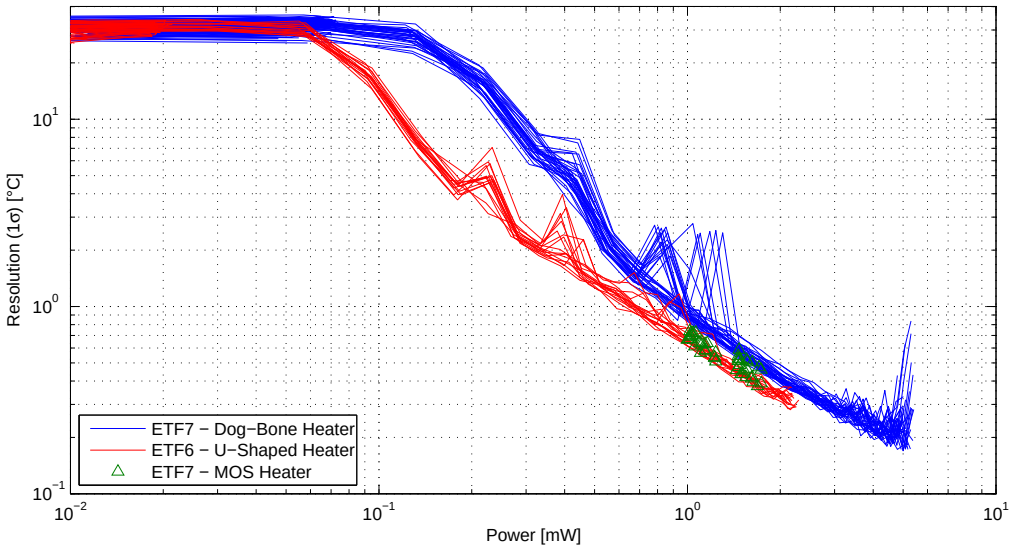

Figure 4.11: Resolution over ETF heater power.

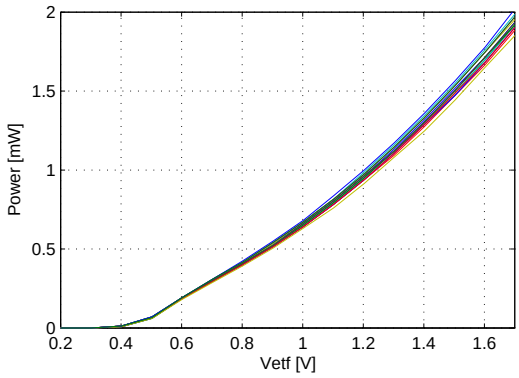

Figure 4.12: ETF6 U-Shaped heater power over heater voltages.

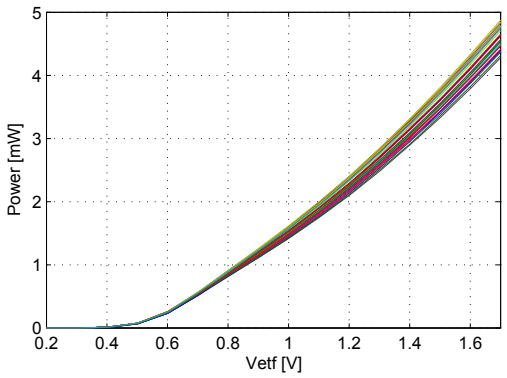

Figure 4.13: ETF7 Dog-Bone heater power over heater voltages.

Table 4.2: Power comparison.

Work	Heater power [mW]	Digital power [mW]	Analog power [mW]
This	1.55	0.53	0.36
[11]	1.8	1.44	0.36
[10]	2.5	0.1	0.5

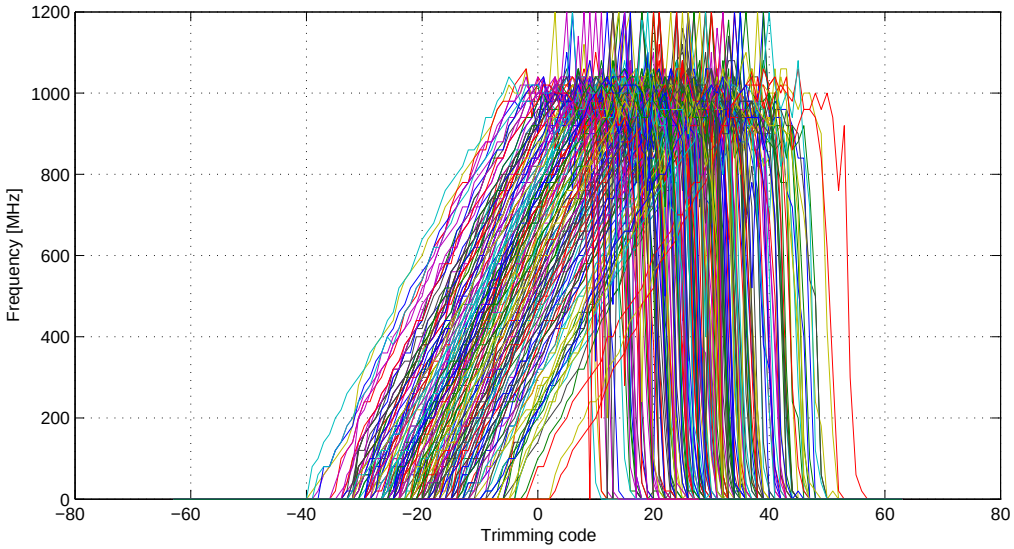

Figure 4.14: CCO Frequency over trimming codes at 25 °C.

level as the MOSFET heaters. To find a good working regime for the resistive heaters, figures 4.11, 4.12 and 4.13 have been made. The dog-bone heater has a lower signal-to-noise ratio than either the U-shaped heater or the MOS heater, this can be explained by the reduced efficiency (see figure 4.10).

Finally we can report a power comparison, see table 4.2. Since the sensor of [10] does not have any significant digital circuitry and digital power consumption, its power consumption has not been reported per domain. The analog and digital power of that publication reported here have been estimated. The digital power of [11] compared to this work are probably due to the problems with the tripler. The simulated power consumption of [11] was 1.27 mW, which is lower than the measured power consumption. The measured and simulated power consumption of this work do match quite well, (with a 0.03 mW difference).

4.3.2. CCO tuning frequency

Due to offset in the Gm stage and process-spread and mismatch in the CCO, the nominal frequency of the CCO will spread. The system employs an IDAC to trim-out these errors. In figure 4.14 the CCO frequency is plot over the different IDAC trimming codes. From this plot it is clear that the counter can handle frequencies up to 1 GHz. After this point the counter starts to fail.

When trimming all CCOs to 500 MHz, and then plotting the frequency behavior over deviations of the IDAC trimming code (See figures 4.15, 4.17 and 4.19) we can see that there is

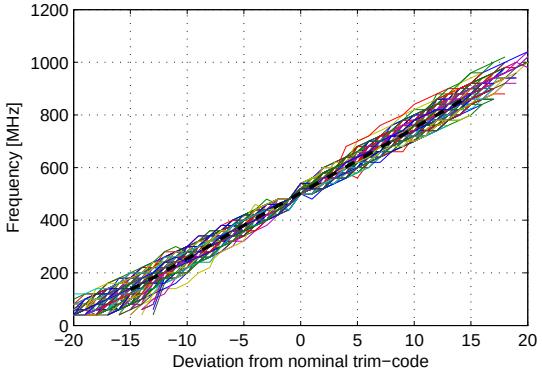

Figure 4.15: Frequency tuning curves at -35 °C.

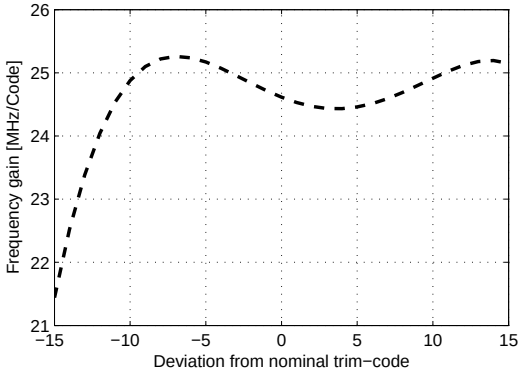

Figure 4.16: Frequency tuning, master-curve fitted, gain at -35 °C.

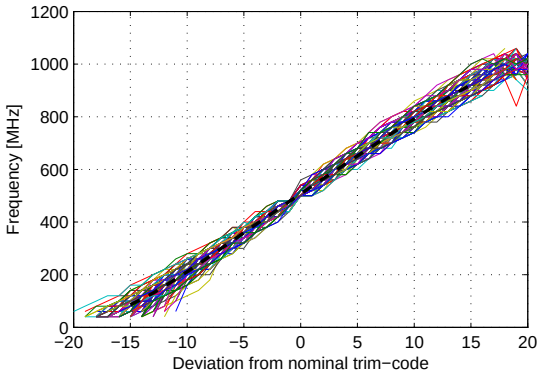

Figure 4.17: Frequency tuning curves at 25 °C.

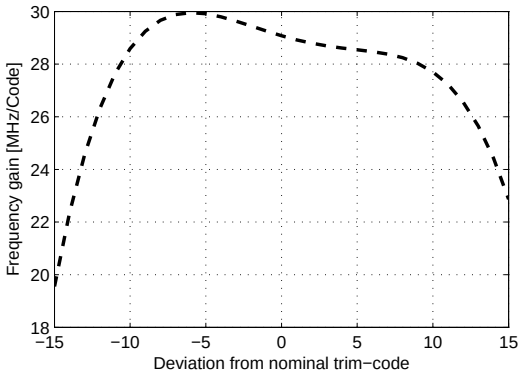

Figure 4.18: Frequency tuning, master-curve fitted, gain at 25 °C.

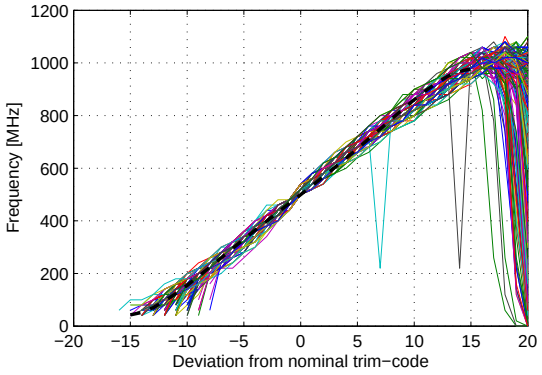

Figure 4.19: Frequency tuning curves at 125 °C.

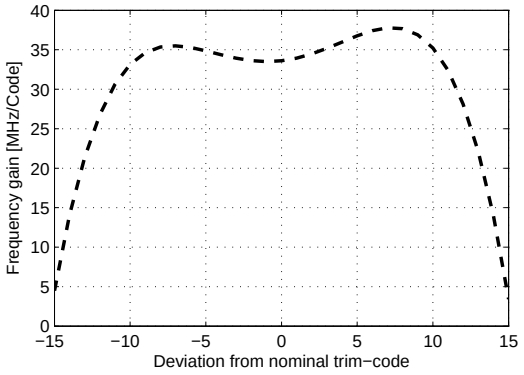

Figure 4.20: Frequency tuning, master-curve fitted, gain at 125 °C.

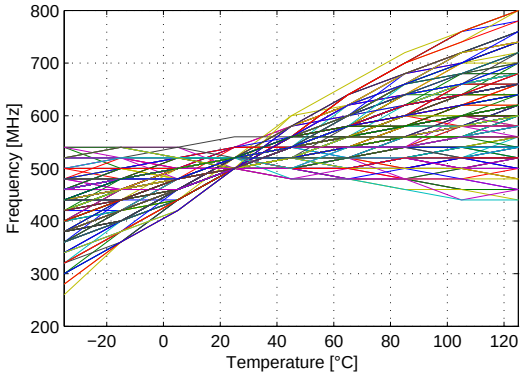

Figure 4.21: VCO Temperature sensitivity, for VCO trimmed at 25 °C to 500 MHz.

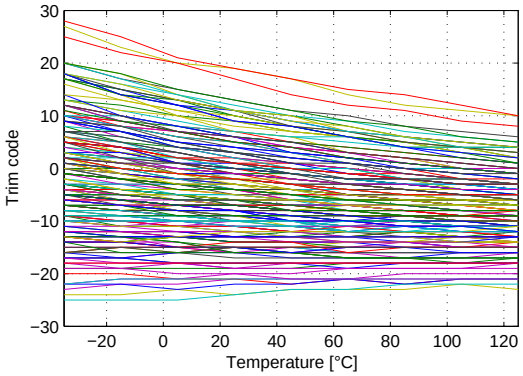

Figure 4.22: Trimming codes for VCO's at different temperatures for VCO trimmed to 500 MHz.

almost no gain mismatch. For different temperatures however we do see different gains (See figures 4.16, 4.18 and 4.20), which is not necessarily only due to temperature dependencies in the CCO, but can also be explained by the constant-Gm biasing employed for the system and hence in the IDAC. The glitches which can be seen in figure 4.19 and 4.40 are due to a bug in the FPGA programming, which is caused by meta-stability in the trimming algorithm, and has been resolved in later versions of the software.

Finally we look at the temperature dependency of the CCO frequency, and 4.22. By trimming the CCO at 25°C, and using that trimming code for all temperatures, see figure 4.21, we can see a very linear temperature dependency in the frequency, which corresponds to the constant-Gm biasing. By trimming the CCO at all temperatures and plotting the trimming code at these temperatures, see figure 4.22, we can see that the trimming codes are well within the boundaries of -64 to 64.

4.3.3. Phase Noise and Stability

The noise is measured at room temperature by taking 10 000 conversions, each of these conversions is transformed into a temperature value, then for each pixel the standard deviation is determined. This 1σ resolution value is then averaged for all pixels to come at the final result.

Figure 4.23 shows the power spectral density of a typical bitstream. For a 500 Hz bandwidth this corresponds to a resolution of 0.47°C, which is fully limited by white spectral noise. The peaking in the spectrum is explained by the delay in the PDΣΔM loop. Especially the new counter has a longer delay, while the previous counter employed carry-bypass circuitry to make it fast in a synchronous way, the new counter trades in delay for speed. However this is no problem since the resolution is still limited by white spectral noise and not quantization noise.

The lower resolution of this design, compared to the analog readout [10], is mainly due to the lower power burned in the heater. However, while burning a similar amount of energy in the heater, the resolution was still slightly worse than that of the analog readout; 0.29°C compared to 0.21°C. To find the reason for this difference a more thorough analysis on the VCO noise has been done in appendix A. It appears that the increased noise floor is due to the quantization of time in the counter. Furthermore this noise floor is higher than expected during the previous design, mainly because this noise correlates between sequential PDΣΔM periods.

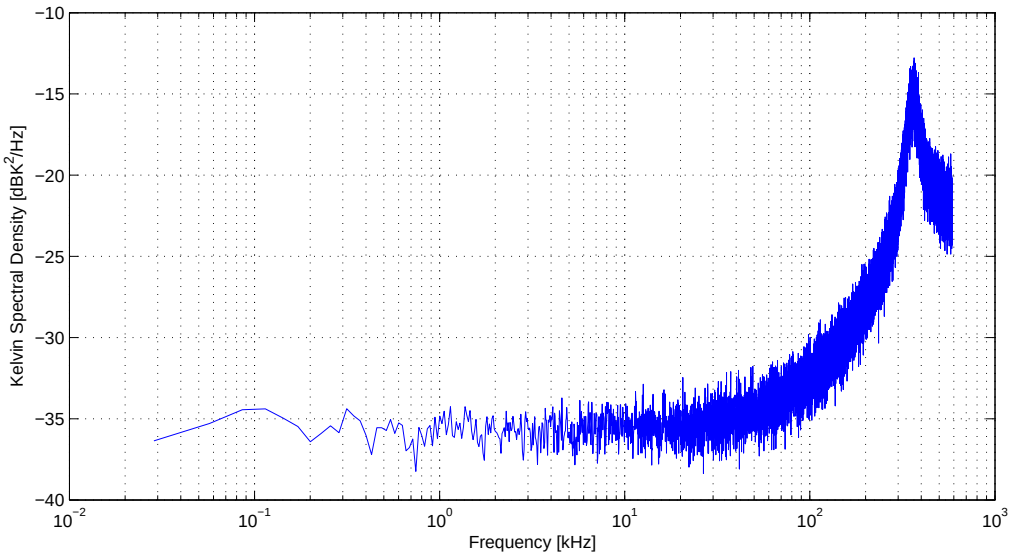

Figure 4.23: FFT of the PDΣΔMs bit-stream for 40 960 samples averaged over 25 FFT's.

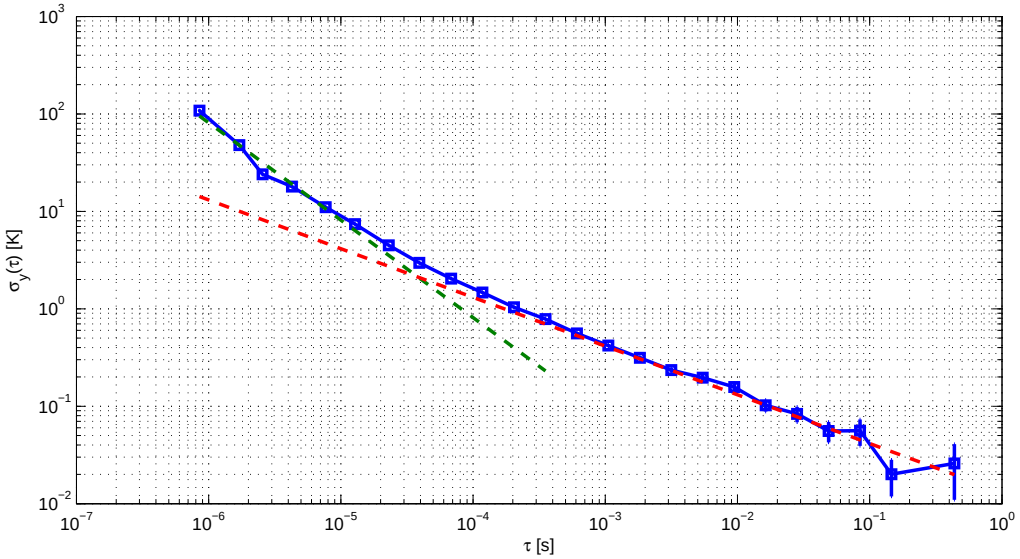

Figure 4.24: Allan Deviation in Kelvins.

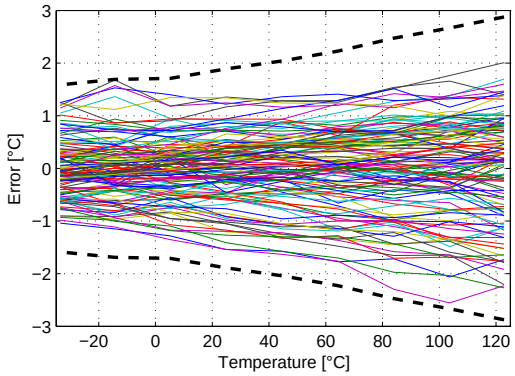

Figure 4.25: Error plot of square ETF with MOSFET heater with Low-Power Counter.

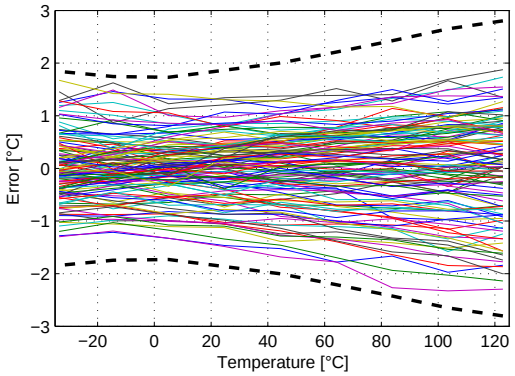

Figure 4.26: Error plot of square ETF with MOSFET heater with Low-Power Counter with fine conversion.

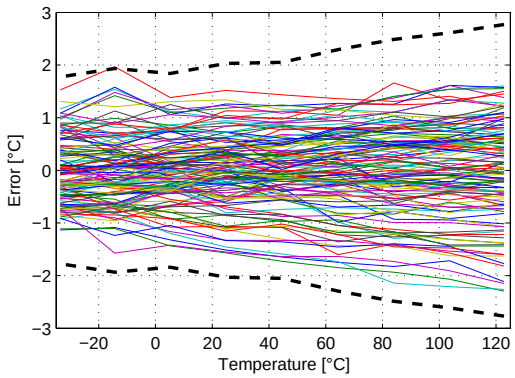

Figure 4.27: Error plot of square ETF with MOSFET heater with conventional counter.

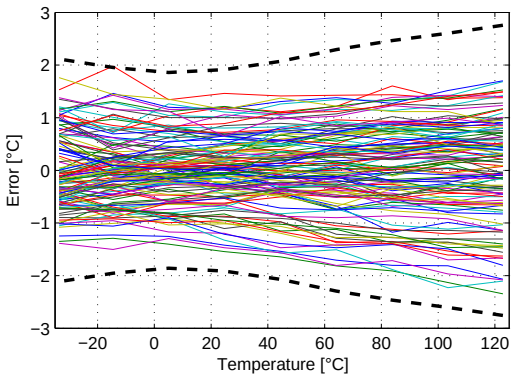

Figure 4.28: Error plot of square ETF with MOSFET heater with conventional counter with fine conversion.

Finally a stability test is done using the Allan variance, see figure 4.24. The Allan deviation is done on the bit-stream from the noise measurement. In figure 4.24 we can clearly see the corner at 0.4µs, where the thermal noise takes over from the quantization noise. Also the peak from the FFT (figure 4.23) can be seen in the Allan deviation plot by a small bump. For the lower frequencies a longer conversion of a few hours should be done, right now there is not enough data to make any meaningful conclusions on the accuracy limit. The most important conclusion to take-away is that for the approximately 0.1 s that each pixel is measured the remaining instability is below 0.1 K.

4.3.4. Accuracy

Figures 4.25 and 4.27 show the spread of the two different kind of pixels with the MOSFET heaters. Since both are MOS heaters they are very similar. Also the spread is plotted for the fine conversion for both the pixels, see figure 4.26 and 4.28. For the full temperature range the 3σ inaccuracy is 2.9°C.

Figures 4.29, 4.30, 4.31 and 4.32 show the inaccuracy of the ETF with the Dog-Bone heater. The errors at the high temperatures in figures 4.30 and 4.32 should be neglected since the PDΣΔM is in its dead-zone for these measurements, which is why the coarse conversion does

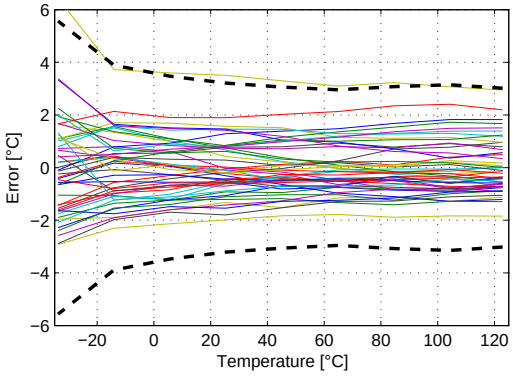

Figure 4.29: Error plot of square ETF with Dog-Bone heater with Low-Power Counter.

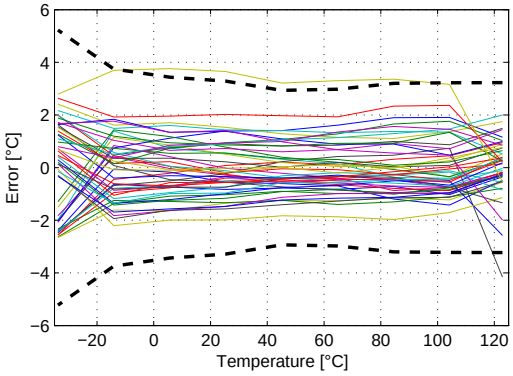

Figure 4.30: Error plot of square ETF with Dog-Bone heater with Low-Power Counter with fine conversion.

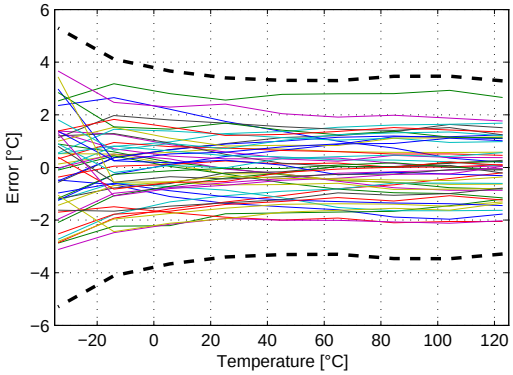

Figure 4.31: Error plot of square ETF with Dog-Bone heater with conventional counter.

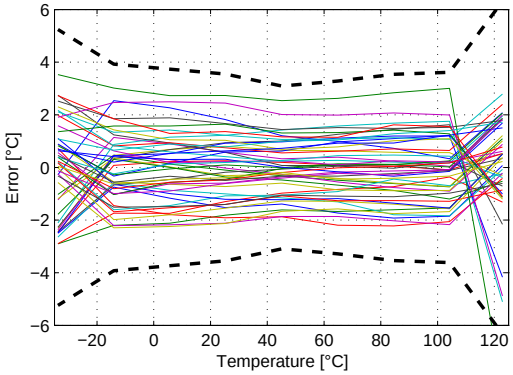

Figure 4.32: Error plot of square ETF with Dog-Bone heater with conventional counter with fine conversion.

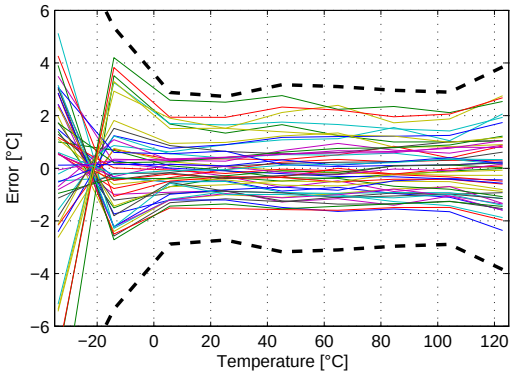

Figure 4.33: Error plot of rotund ETF with U-Shaped heater with conventional counter.

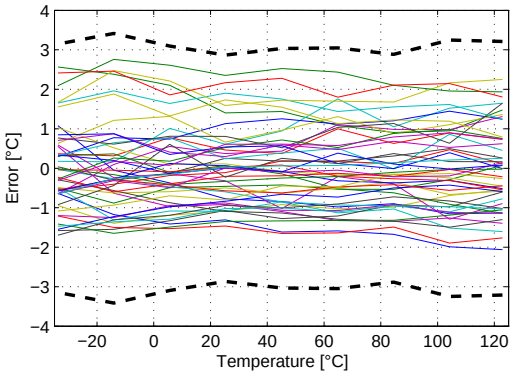

Figure 4.34: Error plot of rotund ETF with U-Shaped heater with conventional counter with fine conversion.

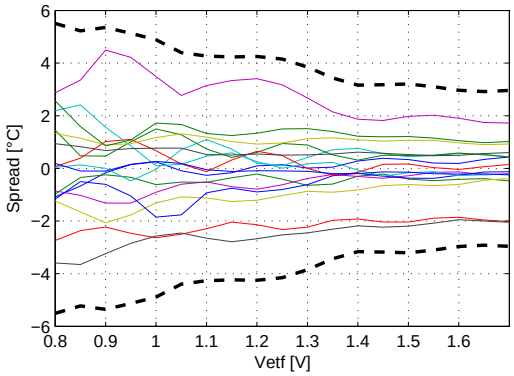

Figure 4.35: Accuracy (3σ) of ETF6 U-Shaped heater over heater voltages.

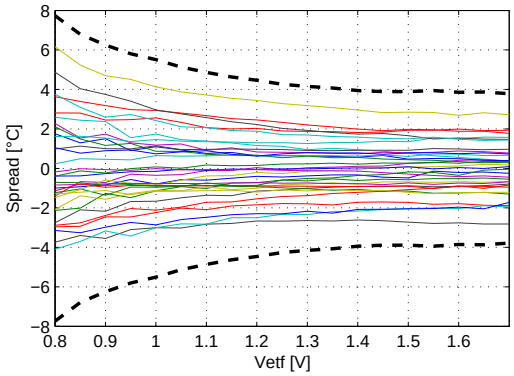

Figure 4.36: Accuracy (3σ) of ETF7 Dog-Bone heater over heater voltage.

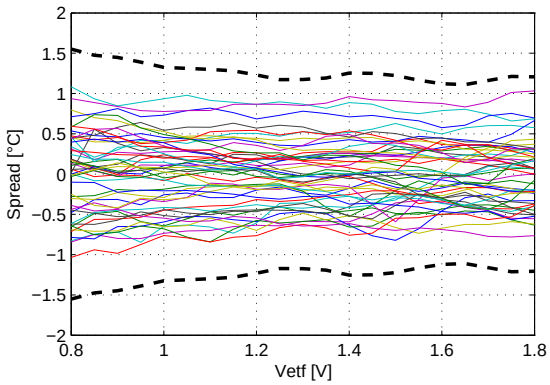

Figure 4.37: Accuracy (3σ) of MOS heater over heater voltages.

not converge to a correct fine value for every conversion. Figures 4.33 and 4.34 show the inaccuracy of the ETF with the U-shaped heater. For this ETF the phase readout at the lowest temperatures is outside of the phase range of the PDΣΔM, so at -35°C these results should be neglected.

Clearly the inaccuracy for the resistive heaters is much worse than that of the MOSFET heater, this is especially the case for the Dog-Bone heater. This can be explained by the inherent inefficiency of the heaters, see figure 4.10, which is also much worse for the Dog-Bone heater. This means that the heat produced in the ETF varies, which varies the self-heating effect, but more importantly it varies the amount of heat produced in the switches. Since the switches are near the thermopile's cold junction, this causes thermal cross-talk. By measuring the spread of the sensor readout over the heater voltage we can see this effect, see figures 4.35 and 4.36. Compared to the MOSFET heater, see figure 4.37, where only the self-heating effect can be seen, which is much smaller.

The measured temperature sensors still exhibit a larger temperature spread than the analog readout, i.e. 2.9°C compared to 2.4°C. The reason for this difference has been found while checking the phase-calibration measured delay for different CCO tuning frequencies. It was found that there is a dependency between the two, which means that different trimming frequencies will show different measured delays. The analysis on the VCO based PDΣΔM shows

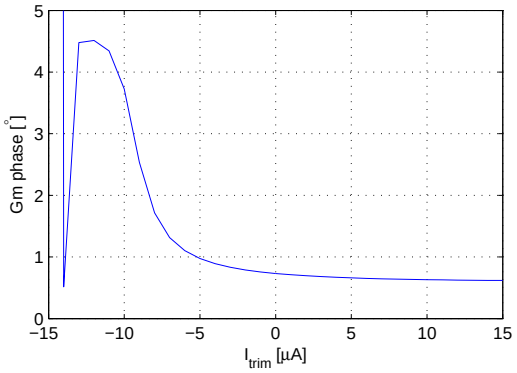

Figure 4.38: Gm-stage phase delay at 1.1719 MHz over trimming current.

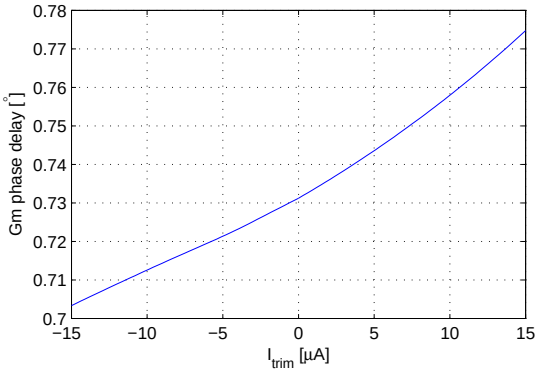

Figure 4.39: Gm-stage phase delay at 1.1719 MHz over trimming current, with folded-cascode compensation current.

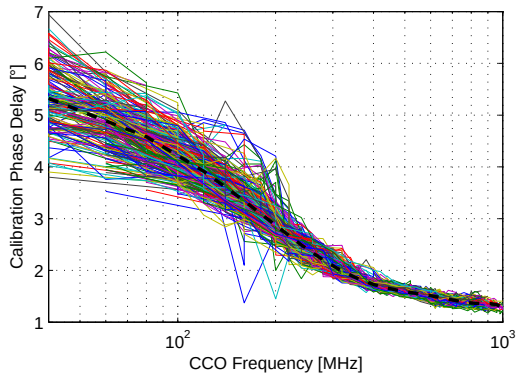

Figure 4.40: Calibration delay over CCO trimmed frequency.

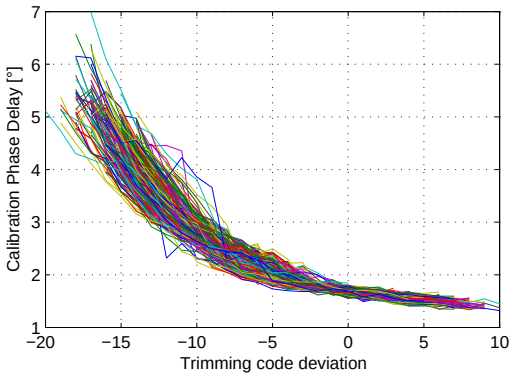

Figure 4.41: Calibration delay over trimming code deviation.

that such a frequency dependency should be canceled out. After some deliberation the cause of this dependency has been found; the trimming current is directly taken from the the input-pair current and cascode current, thereby changing the phase delay of the Gm stage. This effect has been simulated, see figure 4.38. The major cause for the large phase shift is that, when trimming down the cascode will get less and less current, until in simulation the current completely stops. To find out which effect is more dominant, the current change in the input pair or in the cascode, figure 4.39 has been made, where current sources were put around the folded cascode to always have the same current going through the folded cascode. These figures suggest that by mismatch of the Gm and CCO the cascode will cause a spread in the phase delay. This is proven by comparing figure 4.38 with figure 4.41, which is the measured delay for different rimming currents. When keeping in mind that at room temperature the unit current in the IDAC is around 0.5µA, we see a perfect match in measured and simulated delay.

From Monte-Carlo simulations it was found that the 3σ inaccuracy of the trimming current is 11.1µA, 10.6µA and 12.8µA for -55°C, 27°C and 125°C respectively. Taking only the effect of the input pair into account, and not that of the folded cascode (see figure 4.39), this would give rise to a 3σ of about 0.19°C. This suggests that by putting the trimming current inside the Gm stage, next to the input pair, you get an extra spread of about 0.19°C.

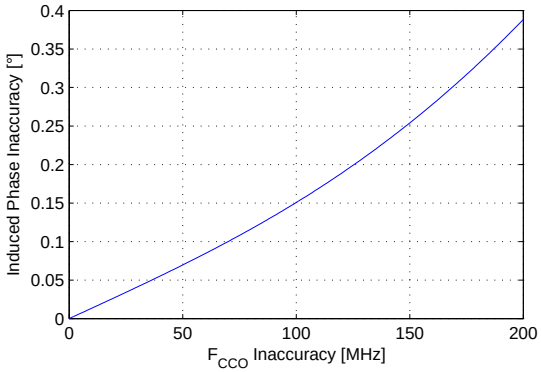

Figure 4.42: Induced phase inaccuracy in readout due to inaccuracy in Fcco trimming.

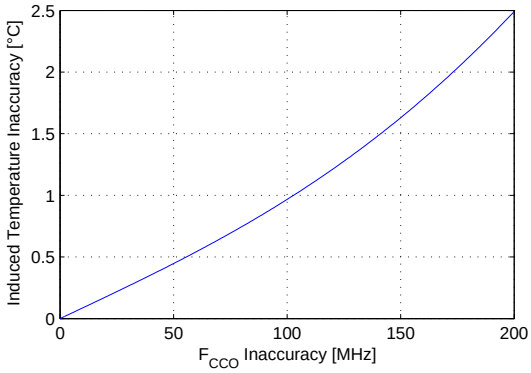

Figure 4.43: Induced temperature inaccuracy in readout due to inaccuracy in Fcco (for the MOS heater at 25 °C).

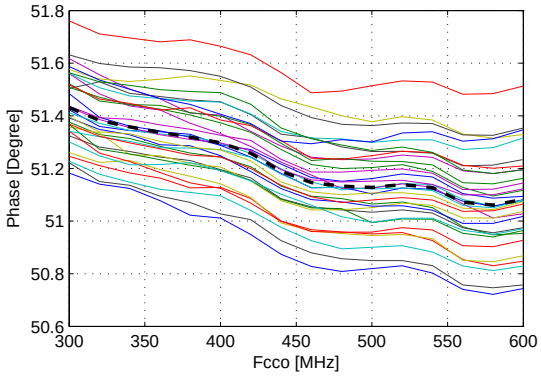

Figure 4.44: Phases of ETF7 Dog-Bone heater for different Fcco frequencies.

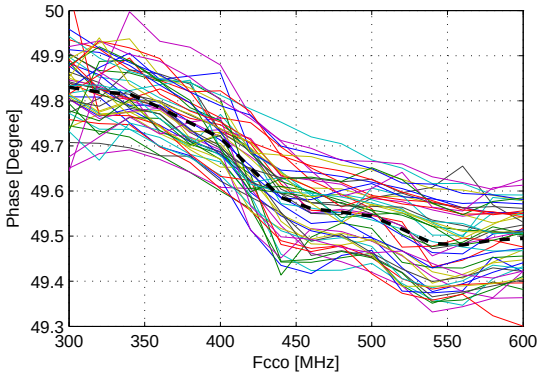

Figure 4.45: Phases of ETF7 MOS heater for different Fcco frequencies.

From the frequency sensitive delay we can also determine, to first order, what kind of induced inaccuracy/noise we get because of CCO frequency sensitivity in the readout. For the 25 °C case this is plotted in figures 4.42 and 4.43. Due to stochastic errors in the trimming process a pixel might be trimmed to a slightly different value than previous conversion, thereby having a different CCO frequency and a different delay, therefore this stochastic process adds noise to the temperature measurement. Figure 4.44 and 4.45 show that a normal ETF phase measurement is also sensitive to the trimming current, and thus the CCO frequency.

4.3.5. Phase-calibration

The phase-calibration did not work to remove or reduce the spread in the temperature readout. The problem is that the phase-calibration delay did not correlate very well with the electrical delay which is present when measuring the ETF phase. This can be seen from figure 4.46, where the spread of the phase-calibration measurement is compared to the spread of the ETF phase measurement.

The reason why the correlation is poor is because of the changing Gm-stage phase response. The phase-calibration is pulled through a single ETF arm at a time, thereby pulling the effective common-mode down. Simulations show that the Gm response changes depending on the phase calibration current strength setting, see figures 4.47 and 4.48. The limited output

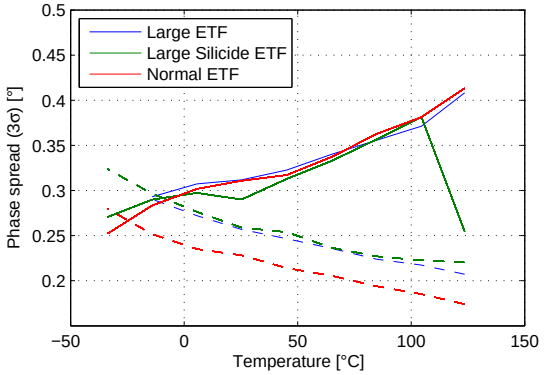

Figure 4.46: Measured spread of phase calibration (dotted lines) and normal measurement (solid lines).

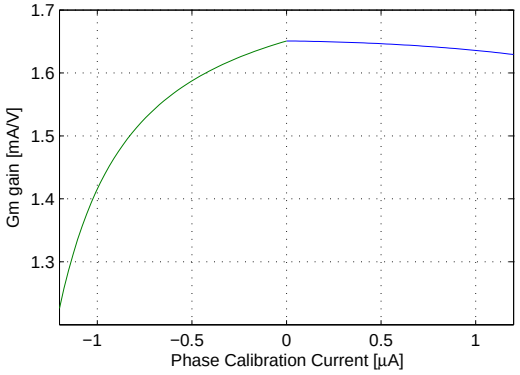

Figure 4.47: Gm gain at 1.1719 MHz for different phase-calibration current driving strengths.

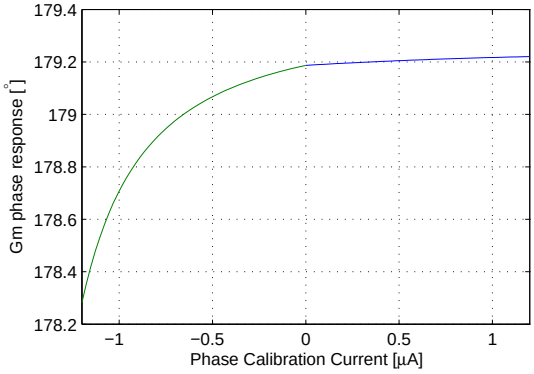

Figure 4.48: Gm phase response at 1.1719 MHz for different phase-calibration current driving strengths.

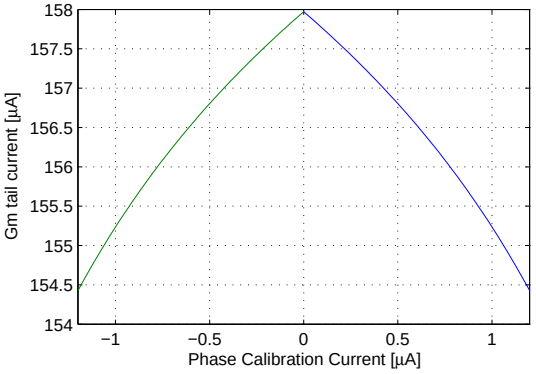

Figure 4.49: Simulated Gm tail current for different phase-calibration current driving strengths.

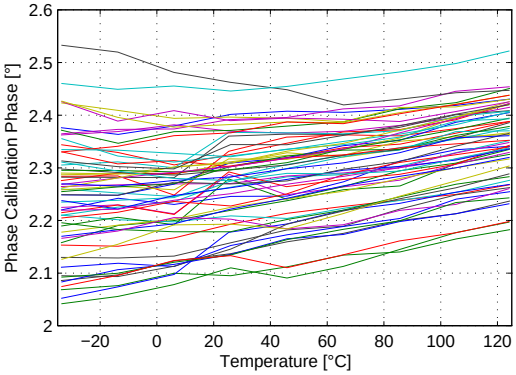

Figure 4.50: Phase-calibration for small phase-cal amplitude.

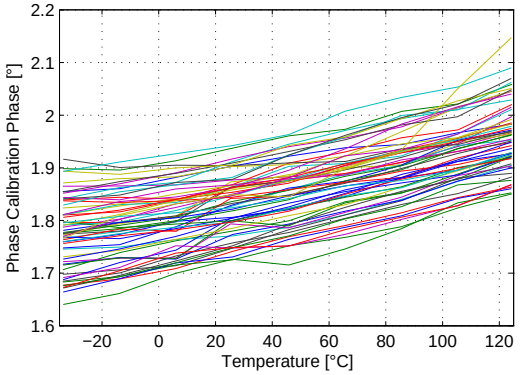

Figure 4.51: Phase-calibration for large phase-cal amplitude.

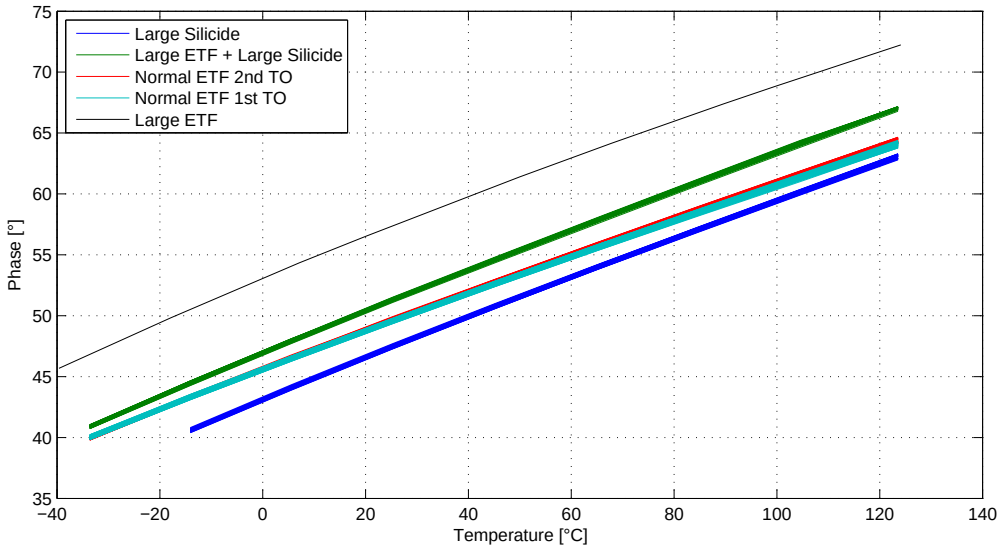

Figure 4.52: ETF master-curves comparison for 2nd TO.

impedance of the tail current source causes some change in tail current, see figure 4.49, while the bias voltage of the tail current source remains the same. This effect can also be shown in measurement, see figures 4.50 and 4.51, where phase calibration has been done using two different phase calibration current strength settings. A definite shift between the two is clearly visible, making the phase calibration more dependent on the phase calibration strength setting than on the electrical delay which is present during normal ETF phase measurements.

4.3.6. Batch-Batch measurements

The batch to batch measurement, which also has the large silicided ETF version, has been measured. Figure 4.52 shows the master curves of the different ETF versions (The data from the Large ETF is taken from [10]). The normal, which is to say the small silicided area ETF, of both batches have a shift between their master-curves of a maximum of 1.3°C, which is smaller than their 3σ inaccuracy. The larger silicide versions have a shifted master-curve with respect to their smaller silicided counter parts. This is because the effective junction is not the effective middle of the silicided area, but is closer to where the most heat is, i.e. the heater. The reason why the large ETF shifts more between the small and large silicided version is because the cold-junctions have less effect on the ETF phase.

The most important information from the large silicided ETFs is their inaccuracy, see figure 4.53. In this figure the 3σ lines for different sensor types have been drawn for easy comparison. The large silicided ETFs seem to have made the readout a bit more accurate; the difference, however, is very small. More importantly the large ETF doesn't seem to have much better accuracy than the smaller ETFs, which further proves that the increased inaccuracy of presented temperature sensors, over the analog version [10], is mainly due to readout errors.

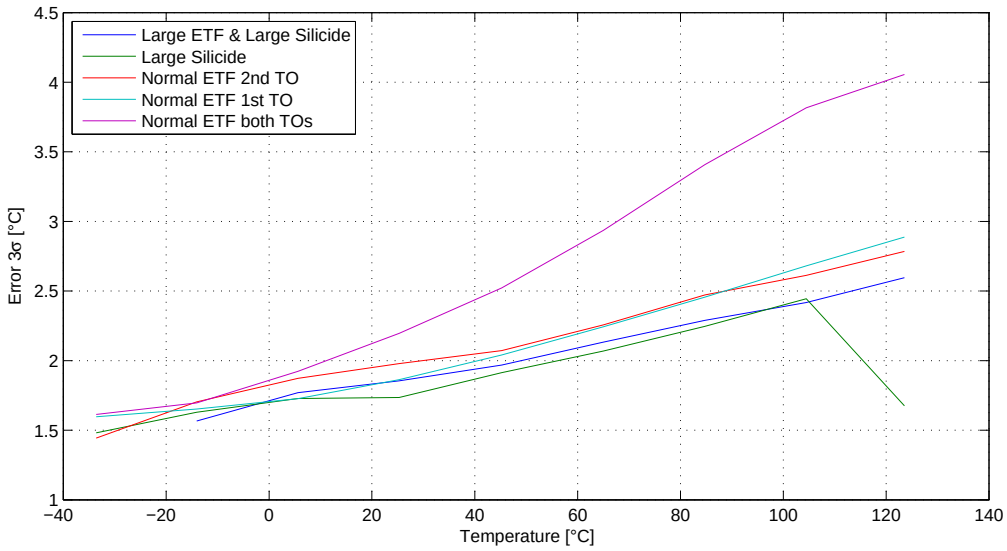

Figure 4.53: Batch-batch accuracy comparison.

5

Conclusion

Table 5.1: Performance Summary and Comparison

	This work	[11]	[10]	[15]	[16]	[6]	[17]
Type	TD	TD	TD	BJT	BJT	MOS	BJT
Process [nm]	160	160	160	32	32	65	14
Area [mm²]	0.0028	0.0046	0.008	0.02	0.02	0.008	0.0087
Inaccuracy, No Trim [°C]	2.9 (3σ)	6.5 (3σ)	2.4 (3σ)	-	5 (max)	-	-
1-pt. Trim [°C]	1.2 (3σ)	1.5 (3σ)	0.65 (3σ)	4.5 (3σ)	-	1.5 (max)	5 (3σ)
2-pt. Trim [°C]	-	-	-	1.2 (3σ)	-	-	2.1 (3σ)
Temp. Range [°C]	-35 — 125	-10 — 125	-40 — 125	20 — 100	-10 — 110	0 — 110	0 — 100
Resolution [°C]	0.47	0.6	0.21	0.19	0.15	0.94	0.5
Speed [kS/s]	1	0.9	1	2	1.2	469	50
Supply Sensitivity [°C/V]	6.5	-	1.3	0.7	-	LDO Reg.	58
Power [mW]	2.4	3.6	3.1	3.8	1.6	0.5	1.1

The smallest smart temperature sensor has been proposed. It achieves close to the analog performance of [10], while using the highly digital VCO based readout presented by [11]. The size of the sensor has been reduced by 42%, while achieving better accuracy and resolution for lower power. Compared to other temperature sensors targeted at thermal management, the sensor is significantly smaller, even though most of them were made in much smaller technologies. The accuracy of the proposed sensor is better than the other type of sensors, both for the no-trim and single point trim, while the other type of sensors also are reported for smaller temperature ranges.

5.1. Future Work

There are two issues in this work which can be solved. The first is the trimming current which, because of its placement in the Gm stage, changes the bandwidth and therefore phase response of the Gm stage. Part of this problem is also the folded cascode, which has a similar bandwidth as the input-pair, but for much smaller currents. Therefore even small current differences in the

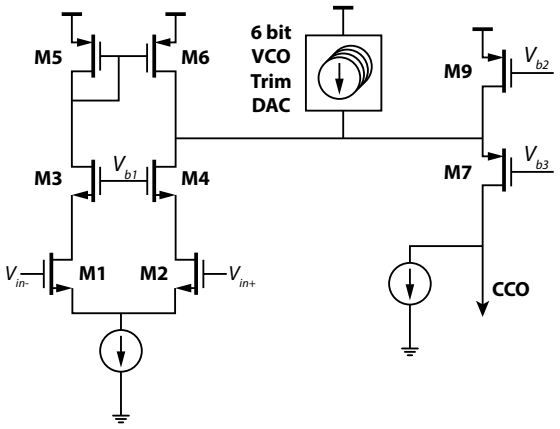

Figure 5.1: Improvement for trimming/Fcco dependency in readout.

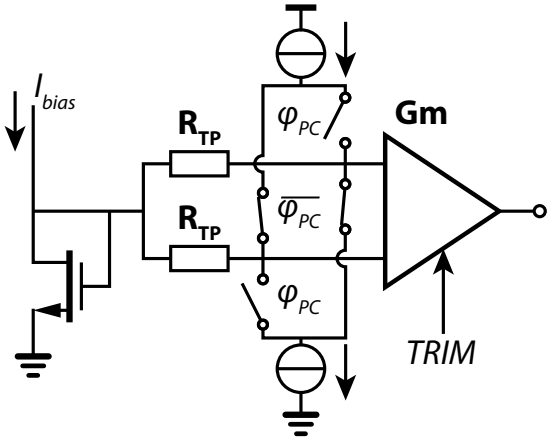

Figure 5.2: Improvement for phase calibration current strength in readout.

cascode will have a measurable effect on the bandwidth. To solve this we apply a bleed current to the cascode, which will give us a smaller sensitivity to the trimming current and therefore better accuracies. A possible solution is shown in figure 5.1. It has the advantage of the IDAC being cascoded, while having an extra current source to bleed the extra cascode current.

The next problem is the phase calibration not correlating to the electrical delay of the ETF phase measurement. Because of the phase calibration implementation there was a DC shift, causing a different phase response of the Gm stage. By running a current through both thermopile arms, see figure 5.2, this DC shift can be mitigated.

Other changes which can be made include increasing the VCO gain to reduce the discretization noise. And increasing the amount of bits in the counter to allow for a larger heater power. Both of these changes will improve the temperature resolution of the sensor.

The next step in a 160-nm process would be to make the sensor a full stand-alone sensor. This means that bit-stream averaging, phase-calibration, trimming etc. must be done by the sensor itself. Performance wise there is not much left to be gained in a 160-nm implementation. By going to smaller technologies the scaling of TD sensors can be proven, therefore better performance will be achieved. On top of that it also opens up the possibility of introducing some new tricks without increasing the area.

A

Discrete-time noise

Because the measured noise of the temperature sensors was higher than expected, a more thorough analysis on the discretization noise of the PDΣΔM has been done. As can be seen in the following the discrete-time artifacts are indeed responsible for the increased noise floor. The reason why this noise floor is higher than expected from the analysis of the previous design is that noise between PDΣΔM periods correlate. Lowering the noise floor can be done by increasing the ETF voltage to frequency gain.

A.1. Discrete-time artifacts

The VCO and counter based readout is inherently a quantized readout. This quantization already happens at the counter clock input, where the input frequency is discretized on the rising edges. The idea is to have enough timing resolution such that discretization/quantization artifacts are below the detection-limit due to thermal noise.

In the following an analysis on the discrete time/quantization noise is done. In the analysis we will approximate the VCO output signal by its first order harmonic only. Since we are only interested in the quantization noise, no other noise sources are modeled. The VCO output signal is then modeled as:

$$f\left(t, \phi_{ETF}\right) \approx A \cdot \cos\left(2\pi t/T - \phi_{ETF}\right) + f_{nom} \tag{A.1}$$

in which A is the amplitude of the VCO frequency swing, T is the PDΣΔM period, ϕ_{ETF} is the ETF phase, and f_{nom} is the nominal, or dc, frequency.

In the ideal case, for every PDΣΔM period, the change in the counter's state will be the difference between the integration value of both the chopper demodulation phases. The first phase, which starts at the feedback phase ϕ_{FB}, lasts for half a PDΣΔM period. The second phase, where the down signal on the counter is active, is then the second half of the PDΣΔM period:

$$
\begin{aligned}
\Delta N_{ideal} &= \int_{t_{\phi_{DAC}}}^{t_{\phi_{DAC}}+T/2} f\,(t, \phi_{ETF})dt \\
&\quad - \int_{t_{\phi_{DAC}}+T/2}^{t_{\phi_{DAC}}+T} f\,(t, \phi_{ETF})dt \\
&= \left[A \cdot \frac{T}{2\pi} \cdot \sin\left(2\pi t/T - \phi_{ETF}\right) + f_{nom} \cdot t\right]_{t_{\phi_{DAC}}}^{t_{\phi_{DAC}}+T/2} \\
&\quad - \left[A \cdot \frac{T}{2\pi} \cdot \sin\left(2\pi t/T - \phi_{ETF}\right) + f_{nom} \cdot t\right]_{t_{\phi_{DAC}}+T/2}^{t_{\phi_{DAC}}+T} \\
&= A \cdot \frac{T}{2\pi}\Big(\quad \sin\left(\pi - \phi_{ETF} + \phi_{DAC}\right) - \sin\left(2\pi - \phi_{ETF} + \phi_{DAC}\right) \\
&\quad -\sin\left(2\pi - \phi_{ETF} + \phi_{DAC}\right) + \sin\left(\pi - \phi_{ETF} + \phi_{DAC}\right) \quad \Big) \\
&\quad + f_{nom} \cdot \left(T_{\phi_{DAC}} + T/2 - T_{\phi_{DAC}} - T_{\phi_{DAC}} - T + T_{\phi_{DAC}} + T/2\right) \\
&= A \cdot \frac{T}{\pi}\Big(\sin\left(\pi - \phi_{ETF} + \phi_{DAC}\right) - \sin\left(2\pi - \phi_{ETF} + \phi_{DAC}\right)\Big) \\
&= -A \cdot \frac{2T}{\pi} \sin\left(\phi_{DAC} - \phi_{ETF}\right)
\end{aligned}
\tag{A.2}
$$

where ϕ_{DAC} is the phase-DAC phase and $t_{\phi_{DAC}}$ is the starting time corresponding to a ϕ_{DAC} phase for a PDΣΔM period.

For a zero phase difference between the feedback and ETF phase this results in zero accumulation. The PDΣΔM will force the accumulated counter value back to zero. As discussed before, for a small phase difference between the feedback and ETF phase the accumulation value can be considered linear with respect to the phase difference:

$$
\Delta N_{ideal} \approx -A \cdot \frac{2T}{\pi} \left(\phi_{DAC} - \phi_{ETF}\right) \tag{A.3}
$$

using this linear model and that:

$$
A = V_{ETF} \cdot G_m \cdot K_{CCO} \tag{A.4}
$$

we come to the linear model in figure A.1.

With a better understanding of the ideal integration value, we can take a look at what happens if we introduce discrete time. The source of the discrete time integration is the fact that the VCO output is only evaluated at rising edges. This is what happens at demodulation, where the demodulating up/down signal, which comes from the phase-DAC, is re-sampled using the VCO signal to avoid any metastability. Any change will take from zero up to a maximum of one VCO period to be detected due to the discrete time effect. This time-delay will depend on the VCO phase, which is uncorrelated with the PDΣΔM phase and therefore exhibits stochastic behavior. On top of that, since the delay ranges from zero up to one VCO period, it is a

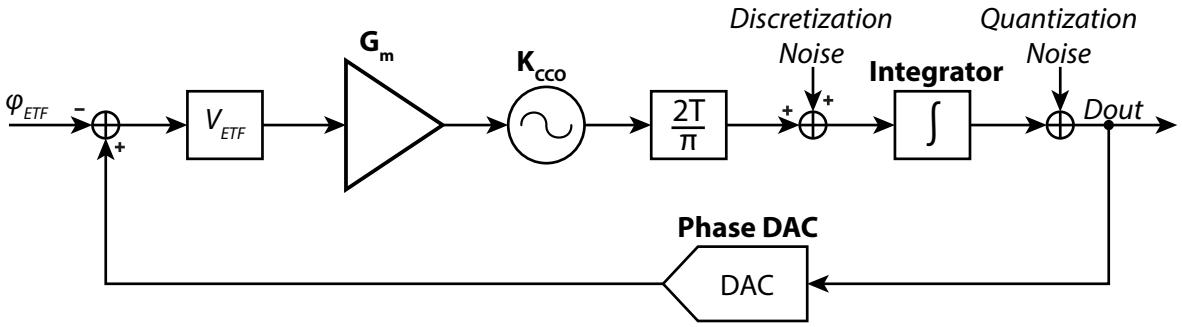

Figure A.1: Simplified linear PDΣΔM noise model.

non-zero mean delay. To model this effect we add μ_{dis} and σ_{dis}, the mean and variance of the discretization artifacts, to the up/down integration periods. Since the time where the up counting ends always, fundamentally, coincides with the time where the down counting begins, the noise sample taken for those two events fully correlate [18], to reflect this these sigmas have the same subscript:

$$\begin{aligned}
\Delta N &= \int_{t_{\phi_{FB}}+\mu_{dis}+\sigma_{dis,1}}^{t_{\phi_{FB}}+T/2+\mu_{dis}+\sigma_{dis,2}} f\left(t, \phi_{ETF}+\mu_{dis}\right)dt \\
&\quad - \int_{t_{\phi_{FB}}+T/2+\mu_{dis}+\sigma_{dis,2}}^{t_{\phi_{FB}}+T+\mu_{dis}+\sigma_{dis,3}} f\left(t, \phi_{ETF}+\mu_{dis}\right)dt \\
&= A\cdot\frac{T}{2\pi}\Big(\sin\left(\pi+\sigma_{\phi_{dis},2}+\phi_{FB}-\phi_{ETF}\right)-\sin\left(2\pi+\sigma_{\phi_{dis},1}+\phi_{FB}-\phi_{ETF}\right) \\
&\quad -\sin\left(2\pi+\sigma_{\phi_{dis},3}+\phi_{FB}-\phi_{ETF}\right)+\sin\left(\pi+\sigma_{\phi_{dis},2}+\phi_{FB}-\phi_{ETF}\right)\Big) \\
&\quad + f_{nom}\cdot\left(\mu_{dis}+\sigma_{dis,2}-\mu_{dis}-\sigma_{dis,1}-\mu_{dis}-\sigma_{dis,3}+\mu_{dis}+\sigma_{dis,2}\right) \\
&= A\cdot\frac{T}{2\pi}\Big(2\cdot\sin\left(\pi+\sigma_{\phi_{dis},2}+\phi_{FB}-\phi_{ETF}\right)-\sin\left(2\pi+\sigma_{\phi_{dis},1}+\phi_{FB}-\phi_{ETF}\right) \\
&\quad -\sin\left(2\pi+\sigma_{\phi_{dis},3}+\phi_{FB}-\phi_{ETF}\right)\Big)+f_{nom}\cdot\left(2\sigma_{dis,2}-\sigma_{dis,1}-\sigma_{dis,3}\right)
\end{aligned} \tag{A.5}$$

This can be simplified further by understanding that $\sigma_{\phi_{dis}}$, the phase uncertainty corresponding to the discretization timing uncertainty for a period of the PDΣΔM, is very small and that for small phase differences between ϕ_{DAC} and ϕ_{ETF} the three sine-wave terms can be approximated to be linear:

$$\Delta N \approx A\cdot\frac{T}{2\pi}\left(-2\cdot\sigma_{\phi_{dis},2}-\sigma_{\phi_{dis},1}-\sigma_{\phi_{dis},3}\right)+2\cdot f_{nom}\cdot\left(2\cdot\sigma_{dis,2}-\sigma_{dis,1}-\sigma_{dis,3}\right) \tag{A.6}$$

To further simplify this, we need to understand the correlation between the start and end uncertainty for every PDΣΔM period. Just like the end of the up period must coincide with the start of the down period, also the end of the down period must coincide with the start of the up period. If we now take a n-period conversion time and assume n to be very small we can approximate the timing uncertainty for such n-period conversion by:

$$\begin{aligned}\sigma_{tot} &= \sigma_{conv,1} + \sigma_{conv,2} + \ldots + \sigma_{conv,n} \\ &= -\sigma_{dis,1} + 2\sigma_{dis,2} - \sigma_{dis,3} - \sigma_{dis,3} + 2\sigma_{dis,4} - \sigma_{dis,5} - \ldots \\ &\quad - \sigma_{dis,2n-1} + 2\sigma_{dis,2n} - \sigma_{dis,2n+1} \\ &\approx \sqrt{n}\left(2\sigma_{dis,even} - 2\sigma_{dis,odd}\right)\end{aligned} \tag{A.7}$$

which, if the noise level $\sigma_{dis,even}$ is equal to $\sigma_{dis,odd}$, simplifies to:

$$\sigma_{tot} \approx 2\sqrt{2 \cdot n} \cdot \sigma_{dis} \tag{A.8}$$

Using the trick in (A.7) and the property (A.9) we can further simply (A.6), see (A.10):

$$\sigma_{\phi} = \frac{2\pi}{T} \cdot \sigma_t \tag{A.9}$$

$$\begin{aligned}\sigma_N &\approx A \cdot \frac{T}{2\pi}\left(-2 \cdot \sigma_{\phi_{dis},2} - \sigma_{\phi_{dis},1} - \sigma_{\phi_{dis},3}\right) + 2 \cdot f_{nom} \cdot \left(2 \cdot \sigma_{dis,2} - \sigma_{dis,1} - \sigma_{dis,3}\right) \\ &\approx A \cdot \frac{T}{\pi}\left(-\sigma_{\phi_{dis},even} - \sigma_{\phi_{dis},odd}\right) + 2 \cdot f_{nom} \cdot \left(\sigma_{dis,even} - \sigma_{dis,odd}\right) \\ &\approx 2 \cdot f_{nom} \cdot \left(\sigma_{dis,even} - \sigma_{dis,odd}\right) - 2 \cdot A \cdot \left(\sigma_{dis,even} + \sigma_{dis,odd}\right) \\ &\approx 2 \cdot \sigma_{dis,even} \cdot (f_{nom} - A) - 2 \cdot \sigma_{dis,odd} \cdot (f_{nom} + A)\end{aligned} \tag{A.10}$$

Since the up/down chopping signals switch at/around the highest and lowest frequency point of the VCO output it is not very surprising that we get a term around $f_{nom} + A$ and $f_{nom} - A$. These noise sources are the approximated amount of noise which is integrated during every period of the PDΣΔM. Therefore, in the linear model, we can model this as a noise source at the input of the integrator, which is coincidentally at the input of the PDΣΔM (see figure A.1). Since the VCO period is completely uncorrelated with the DAC phase, the discretization noise can be modeled by a uniform distribution with a distribution length of one period of the VCO frequency (remember that the mean shift was modeled by a separate parameter which dropped out), moreover since the VCO frequency, and hence the period, depends on the demodulating phase, we get different values for $\sigma_{dis,even}$ and $\sigma_{dis,odd}$:

$$\sigma_{dis,even} = \frac{1}{\sqrt{12}} \cdot \frac{1}{f_{nom} - A} \tag{A.11}$$

$$\sigma_{dis,odd} = \frac{1}{\sqrt{12}} \cdot \frac{1}{f_{nom} + A} \tag{A.12}$$

After Filling in these value we get the following noise source at the input of the integrator:

$$\begin{aligned} \sigma_N &\approx \frac{2}{\sqrt{12}} \cdot \frac{f_{nom} - A}{f_{nom} - A} + \frac{2}{\sqrt{12}} \cdot \frac{f_{nom} + A}{f_{nom} + A} \\ &\approx \frac{2\sqrt{2}}{\sqrt{12}} \approx \frac{2}{\sqrt{6}} \end{aligned} \tag{A.13}$$

This might seem very strange; a unit-less independent noise source. But the noise is in terms of counter bits, which are unit-less. To refer this noise back to a voltage noise at the ETF, we first find how such an ETF noise voltage will look at the input of the integrator. Since the ETF noise source is at the input of the amplifier we multiply by its gain, then the CCO voltage to frequency gain and then the frequency to counter gain. This is then chopped, but because the amplifier bandwidth is much higher than the chopping frequency, nothing happens to the noise floor [19]:

$$\sigma_{int} = \sigma_{ETF} \cdot G_m \cdot K_{CCO} \cdot \frac{2T}{\pi} \tag{A.14}$$

Finally we put in the requirement that the discretization noise must be smaller than the ETF resistance thermal noise. For that we also need to take into account an integration period of one PDΣΔM period for the ETF thermal noise:

$$\begin{aligned} \sigma_{ETF} &> \frac{\sigma_N}{G_m \cdot K_{CCO} \cdot \frac{2T_P}{\pi}} \\ \sqrt{4kTR \cdot \frac{1}{2T}} &> \frac{\pi}{G_m \cdot K_{CCO} \cdot T_P \cdot \sqrt{6}} \\ G_m \cdot K_{CCO} &> \frac{\pi}{\sqrt{4kTR \cdot T_P \cdot 3}} \end{aligned} \tag{A.15}$$

in which T_P is the PDΣΔM period and T is the temperature in Kelvin. Filling in the numbers, $G_m \cdot K_{CCO}$ must be larger than: 250 MHz/mV. If we take into account other noise sources, like the amplifiers (1/f) noise, which increases the noise floor, then the requirement on the VCO gain will be about 170 MHz/mV.

Bibliography

[1] J. Donald and M. Martonosi, *Techniques for multicore thermal management: Classification and new exploration,* in *ACM SIGARCH Computer Architecture News*, Vol. 34 (IEEE Computer Society, 2006) pp. 78–88.

[2] M. A. P. Pertijs, K. A. A. Makinwa, and J. H. Huijsing, *A CMOS smart temperature sensor with a 3σinaccuracy of ±0.1° C from -55° C to 125° C,* Solid-State Circuits, IEEE Journal of **40**, 2805 (2005).

[3] D. Shim, H. Jeong, H. Lee, C. Rhee, D.-K. Jeong, and S. Kim, *A process-variation-tolerant on-chip CMOS thermometer for auto temperature compensated self-refresh of low-power mobile DRAM,* Solid-State Circuits, IEEE Journal of **48**, 2550 (2013).

[4] K. Souri, Y. Chae, K. Makinwa, *et al.*, *A CMOS temperature sensor with a voltage-calibrated inaccuracy of 0.15° C (3σ) from -55° C to 125° C,* IEEE Journal of Solid-State Circuits **48**, 292 (2013).

[5] P. Park, K. Makinwa, D. Ruffieux, *et al.*, *A resistor-based temperature sensor for a real time clock with ±2 ppm frequency stability,* in *European Solid State Circuits Conference (ESSCIRC), ESSCIRC 2014-40th* (IEEE, 2014) pp. 391–394.

[6] S. Hwang, J. Koo, K. Kim, H. Lee, and C. Kim, *A 0.008 mm² 500 µW 469 kS/s Frequency-to-Digital Converter Based CMOS Temperature Sensor With Process Variation Compensation,* Circuits and Systems I, IEEE Transactions on **60**, 2241 (2013).

[7] A. Mahfuzul Islam, J. Shiomi, T. Ishihara, and H. Onodera, *Wide-supply-range all-digital leakage variation sensor for on-chip process and temperature monitoring,* in *Solid-State Circuits Conference (A-SSCC), 2014 IEEE Asian* (IEEE, 2014) pp. 45–48.

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

[9] K. A. A. Makinwa, *Smart temperature sensor survey,* (2015), (Rev: 2015-02-26).

[10] 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), ESSCIRC 2014-40th* (IEEE, 2014) pp. 395–398.

[11] R. Quan, U. Sonmez, F. Sebastiano, and K. A. A. Makinwa, *A 4600 µm² 1.5° C 0.9 kS/s thermal-diffusivity temperature sensor with VCO-based readout,* in *Solid-State Circuits Conference, IEEE International, Digest of Technical Papers.* (IEEE, 2015) pp. 488–489.

[12] S. M. Kashmiri, S. Xia, and K. A. A. Makinwa, *A temperature-to-digital converter based on an optimized electrothermal filter,* Solid-State Circuits, IEEE Journal of **44**, 2026 (2009).

[13] V. Székely and M. Rencz, *A new monolithic temperature sensor: the thermal feedback oscillator,* in *Solid-State Sensors and Actuators, 1995 and Eurosensors IX.. Transducers' 95. The 8th International Conference on*, Vol. 2 (IEEE, 1995) pp. 124–127.

[14] J. R. Bitner, G. Ehrlich, and E. M. Reingold, *Efficient generation of the binary reflected Gray code and its applications,* Communications of the ACM **19**, 517 (1976).

[15] J. S. Shor and K. Luria, *Miniaturized BJT-based thermal sensor for microprocessors in 32- and 22-nm technologies,* Solid-State Circuits, IEEE Journal of **48**, 2860 (2013).

[16] Y. W. Li, H. Lakdawala, A. Raychowdhury, G. Taylor, and K. Soumyanath, *A 1.05 V 1.6 mW 0.45 C 3σ-resolution ΔΣ-based temperature sensor with parasitic-resistance compensation in 32nm CMOS,* in *Solid-State Circuits Conference-Digest of Technical Papers, 2009. ISSCC 2009. IEEE International* (IEEE, 2009) pp. 340–341.

[17] 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,* Solid-State Circuits, IEEE Journal of **50**, 799 (2015).

[18] M. Z. Straayer and M. H. Perrott, *A 12-bit, 10-MHz bandwidth, continuous-time ADC with a 5-bit, 950-MS/s VCO-based quantizer,* Solid-State Circuits, IEEE Journal of **43**, 805 (2008).

[19] 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 **84**, 1584 (1996).