Technische Universiteit Delft

A Cryo-CMOS Voltage Reference for Quantum Computing Applications

Y. Wu

TU Delft

A Cryo-CMOS Voltage Reference for Quantum Computing Applications

by

Y. Wu

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday September 30, 2022 at 9:30 AM.

Student number:	5228867	
Project duration:	September 1, 2021 – September 30, 2022	
Thesis committee:	Dr. F. Sebastiano,	TU Delft, supervisor
	Dr. D. G. Muratore,	TU Delft
	Dr. Q. Fan,	TU Delft

This thesis is confidential and cannot be made public until September 30, 2024.

Abstract

Voltage references are an essential building block for many electronic systems, such as analog-to-digital and digital-to-analog converters, and voltage regulators. Although state-of-the-art voltage references already demonstrated high performance over the standard temperature range (−40 °C to 125 °C), specific applications require an operating temperature far beyond this range. A relevant example is the electronic interface for quantum processors, for which voltage references that can operate down to cryogenic temperatures are needed.

This thesis presents a CMOS-based voltage reference that can work over a wide temperature range from 4 K to 300 K. To achieve a low-drift, high-accuracy reference, it is designed based on the devices' cryogenic behaviour to minimize the nonlinearities of the PTAT- (proportional to absolute temperature) and CTAT (complementary to absolute temperature) voltages and combine them to generate a stable output voltage. Dynamic compensation techniques are employed to reduce the effect of process variations and a switched-capacitor circuit is proposed to minimize the output ripple. The voltage reference is designed to provide a high output voltage of 0.9 V while operating in a nominal supply voltage of only 1.1 V, with the minimum required voltage of 0.96 V. The expected temperature coefficient (TC) and inaccuracy (3σ) over the entire operating temperature range from 4 K to 300 K is expected to be 60 ppm/K and 1 %, respectively after performing a single point trimming. The power consumption at 300 K is 57 µW and is expected to be 17 µW at 4 K .

The targeted specifications of this design are comparable with the state-of-the-art voltage references working at room temperature. To verify the performance at cryogenic temperatures, the design has been taped out in TSMC-40nm technology.

Acknowledgements

First, I would like to thank my supervisor Dr. Fabio Sebastiano for giving me the opportunity to work on this project. I would like to thank him for all the discussions, valuable comments, and encouragement, which helped me a lot throughout the project. Next, I would like to thank my daily supervisor Job for all the support. I really enjoyed all the discussions we had, technical ones, and of course, non-technique ones. I would like to thank him for always being approachable and always helping me in various ways. It was definitely a great experience working with him. Furthermore, I would like to thank CoolGroup for all the help and the fun time we had. Especially a thank to David and Rishabh, for all the discussions and practical/mental support this year. It was a great time working in the same office with them.

Finally, I would like to thank my family for all their love and support.

Y. Wu
Delft, September 2022

Contents

1
Introduction

1.1. Quantum Computing

Over the past years, an increasing amount of attention and resources has been invested in developing quantum computers. Quantum computing is believed to be a great candidate to solve complex computational problems that cannot be solved in a reasonable time by classical computers. In contrast to the traditional classical computers, which use classic bits 0 and 1 for operation, quantum computers use quantum bits (qubits) as the basic computational unit. Qubits can be in a superposition of both 0 and 1 at the same time [1]. Properties such as superposition and entanglement enable quantum computers to perform parallel operations on large amounts of data, opening the door for an exponential speed-up in the number of calculations.

1.2. Cryogenic Electronics

Quantum mechanical effects are usually associated with very small energy scales. In order to observe this behaviour, qubits must be operated at sub-Kelvin temperatures [2], requiring the qubits to be placed inside so-called dilution refrigerators. Furthermore, quantum processors require an electronic interface to control and read out the qubits. Currently, these electronics are usually operated at room temperature. Therefore, wiring is required to connect the qubits inside the dilution refrigerator to the room-temperature equipment. As a result of the large physical size of such dilution refrigerators, the required wiring has a minimum length in the order of 2 m. This long wiring now also comes with nonidealities such as delay, interference, reliability, resistance, etc. Besides, to perform meaningful computations, future quantum computers would require millions of qubits and therefore millions of wires. Due to the physical space and heat load of the wiring, directly wiring all qubits to room temperature equipment is not feasible and limits the scaling capabilities of quantum processors.

A proposal for a scalable quantum processor is shown in figure 1.1. To reduce the amount of required interconnects in the dilution refrigerator, the electronic interface is proposed to be placed spatially close to the quantum processor [3]. If all the read-out and control electronics can operate at the same temperature and can be integrated with the qubits, many possibilities open up for scaling up the number of qubits. However, such control electronics at cryogenic temperatures also require cryogenic electronic technology. Among semiconductor technologies, CMOS technology has several advantages, for example, it has high reliability and can be integrated on a large scale. Moreover, it has been demonstrated to operate down to 30 mK [4]. These features make CMOS a promising solution for implementing future quantum computers.

1.3. Wide Temperature Range Voltage References

Commonly used electronic circuits, such as data converters and voltage regulators require a well-defined reference voltage. Many designs for high-performance voltage references over the standard temperature range (−40 °C to 125 °C) can be found in literature. However, figure 1.2 shows that certain applications require electronics that can work outside this temperature range. For example, planetary exploration missions and deep space probes rely on low-temperature electronics that can work from −20 °C to −229 °C [5]. Quantum proces-

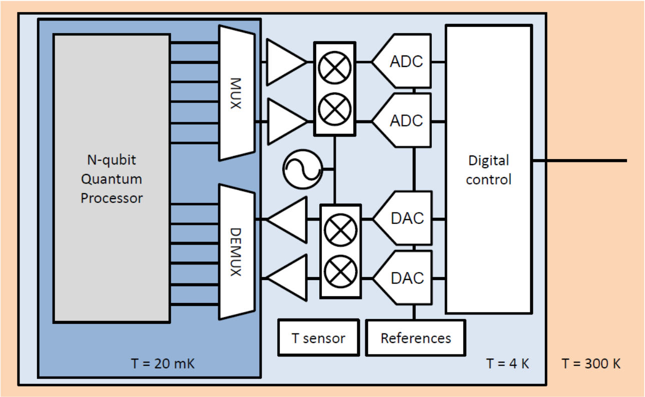


Figure 1.1: Quantum processor proposed in [3]. (Image reproduced from [3])

sors today are also required to work at milli-Kelvin temperatures [2]. For these emerging applications, voltage references that can operate down to the cryogenic temperature are thus required.

This project focuses on designing a wide-temperature range voltage reference for quantum computing applications, although the actual temperature range that the electronics are going to work on is restricted to cryogenic temperatures. Performing standard compensation techniques, such as trimming, is non-trivial and costly if the voltage references would need to be placed inside the dilution fridge before performing the compensation [3]. On the other hand, if it would become possible to perform the calibration at room temperature, then the cost and time required for calibration will be significantly reduced. In addition, since cryo-CMOS design is still in the early stages, extending the knowledge on performance limiting effects for voltage references is of high importance. Extending the temperature range to assess the fundamental limits is therefore also relevant from a scientific point of view.

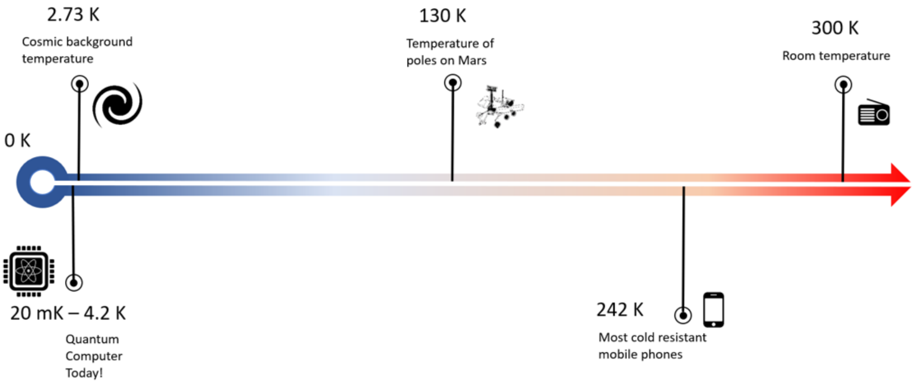

2.73 K
Cosmic background temperature
130 K
Temperature of poles on Mars
300 K
Room temperature
0 K
20 mK – 4.2 K
Quantum Computer Today!
242 K
Most cold resistant mobile phones

Figure 1.2: Applications that require electronics operating outside the standard temperature range.

1.4. Thesis Objective

This project is a follow-up project on the work presented in [6] and [7]. The main objective of this project is to implement voltage references that can operate over a wide temperature from 300 K down to 4 K and to assess to what extent the performance of cryogenic voltage references can compete with the start-of-the-art voltage references over the standard temperature range. The research question for this project can be summarized in the following problem statements:

- How would nonlinear transistor behaviour affect the performance of a voltage reference? Where do these nonlinearities originate from and how do they behave over temperature?
- What is the effect of process variations on the performance of voltage references at cryogenic temperatures? How can they be compensated for?

- What are the bottlenecks for designing a voltage reference that can operate down to cryogenic temperatures? What can be done to overcome these bottlenecks?

One of the main challenges in designing cryo-CMOS and cryogenic voltage references, in general, is the lack of cryogenic CAD-compatible transistor models, while high-performance voltage references would require accurate models to take the second-order effects and process variations into account. For the designs presented, only the standard room temperature models were available. Additional care has thus been taken to ensure high performance and robustness to unexpected changes in device parameters.

1.4.1. Target specifications

An overview of the target specifications is given in table 1.1. The design will be implemented in a bulk 40-nm CMOS process from TSMC, with a nominal supply voltage of 1.1 V. Extended details and a rationale for the derivation of the specifications are discussed in section 2.2.

Table 1.1: Target specifications

Technology	TSMC 40-nm CMOS
Supply	1.1 V
Temperature range	4 K to 300 K
Temperature coefficient (TC)	10 ppm/K
Power	10 μW
Line regulation	10 mV/V
Noise (1Hz to 10Hz)	13.7 μV_{rms}
Inaccuracy (3σ)	1 %
Area	1 mm^2

1.5. Thesis Outline

The thesis is organized as follows: Chapter 2 introduces the working principle of the voltage references and presents the state-of-the-art voltage references at both room temperature and at cryogenic temperatures. This chapter also includes an overview of the design challenges at cryogenic temperatures. Chapter 3 focuses on investigating the nonlinearities presented in the transistor behaviour and the effects of process variations on the reference voltage based on both theory and measurement data. This chapter aims to provide a critical understanding of what is required for designing a cryogenic voltage reference. Chapter 4 presents the system-level design, including the architectural choice and trade-off involved, followed by the transistor-level implementation, simulation results, and layout images. Chapter 5 presents the performance of the design. Finally, the conclusion and future work of the thesis will be presented in Chapter 6.

2

Integrated Voltage References

Voltage references are an essential building block for many electronic systems, such as data converters and voltage regulators. The performance of a voltage reference will directly affect the circuits in which the output of the reference is used [8, 9]. Voltage references should be able to provide a well-defined voltage that is independent of process, voltage, and temperature (PVT-independent). However, it is non-trivial to directly generate a constant voltage over temperature based on a single device parameter. Therefore, additional circuit techniques are usually required to ensure the reference has sufficient temperature stability and also is robust enough against supply and process variations. This chapter aims to provide the background for voltage references, including the working principle, performance metrics, and a state-of-the-art overview. In addition, the rationale behind the targeted specifications of this project (table 1.1) and the design considerations at cryogenic temperatures will be introduced.

2.1. Working Principle

The common approach to generate a reference voltage is based on combining voltages with opposite temperature dependence. For example, adding a voltage that is proportional to absolute temperature (PTAT) to a voltage that is complementary to absolute temperature (CTAT) with a proper scaling factor α results in a voltage that is independent of temperature (V_{ref}) [10, 11].

$$V_{ref} = V_{ctat} + \alpha V_{ptat}. \tag{2.1}$$

This is illustrated in figure 2.1 below.

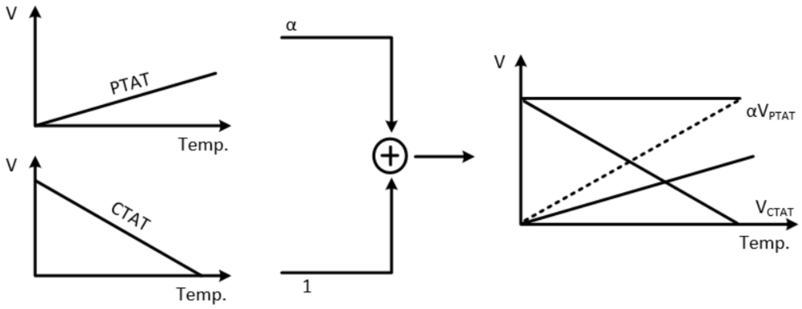


Figure 2.1: Working principle of the voltage references.

2.1.1. BJT-based voltage references

The PTAT- and CTAT voltage can be generated by using BJT transistors as illustrated in figure 2.2. For a forward-biased BJT, its collector current (I_c) is given as

$$I_C = I_S \left[\exp\left(\frac{qV_{BE}}{kT} \right) - 1 \right], \tag{2.2}$$

where I_s is the saturation current, V_{be} the base-emitter voltage, q the electron charge, T the absolute temperature, and k the Boltzmann constant. Since the exponential term in the equation 2.2 is much larger than the -1 term, V_{be} can be obtained by rearranging the equation,

$$V_{be} = \frac{kT}{q} \ln\left(\frac{I_C}{I_S} \right). \tag{2.3}$$

V_{be} is to first order, having a negative temperature coefficient, and therefore can be taken as the CTAT voltage.

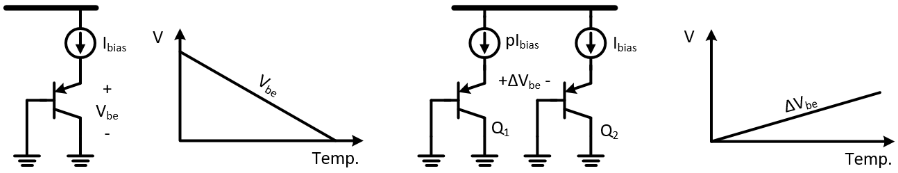


Figure 2.2: Generation of the PTAT- and CTAT voltages using BJTs.

The PTAT voltage can be generated by exploiting the exponential relation between I_C and V_{be}. Assuming two transistors Q_1 and Q_2 are biased with a current density with a ratio of p, then the difference in their V_{be} can be expressed as

$$\Delta V_{be} = V_{be1} - V_{be2} = \frac{kT}{q} \ln(p). \tag{2.4}$$

Adding V_{be} and ΔV_{be} with a proper scaling factor α can result in a voltage V_{ref} that is to first order independent of temperature.

$$V_{ref} = \alpha \cdot \Delta V_{be} + V_{be} = \alpha \cdot \frac{kT}{q} \ln(p) + V_{be}. \tag{2.5}$$

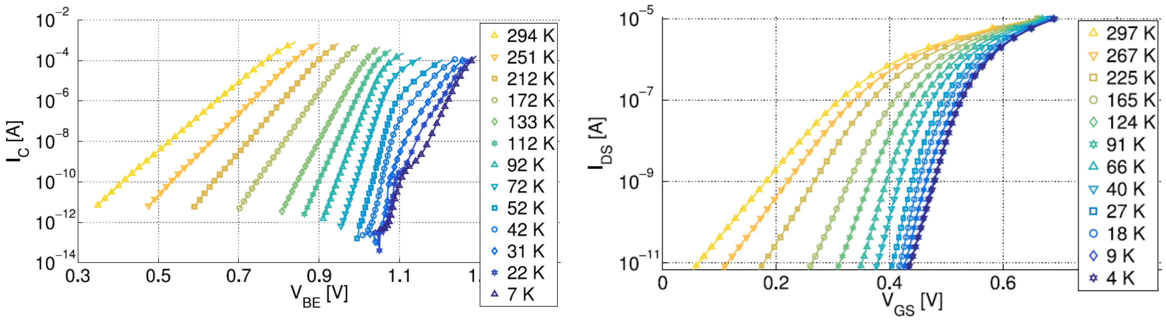
Voltage references that are realized based on this approach are usually referred to as bandgap references as their working principle is fundamentally extracting the bandgap energy of the silicon at 0 K [10–12]. The bandgap voltage at 0 K can be seen as a physical constant and thus is immune to PVT variations. This feature makes it suitable for being a quantity to extract and hence to use as a reference. However, the Si BJT loses its exponential characteristic at cryogenic temperatures due to the reduced current gain and an increased base resistance as shown in 2.3a [13]. Implementing a reference for wide-temperature range operation based on BJTs is therefore not viable.

2.1.2. MOS-based voltage references

Fortunately, MOS transistors in weak inversion show a similar I-V curve as BJT devices, therefore the same concept mentioned above can also be realized by MOS devices. Instead, the threshold voltage of the MOS transistor is extracted, contrary to the bandgap voltage in the case of BJT devices. Moreover, it has already been proven that the exponential behaviour of MOS devices holds down to cryogenic temperatures as shown in figure 2.3b [13, 14], which makes them good candidates for implementing cryogenic voltage reference. The drain-current I_d of MOS devices in weak inversion is an exponential function of its gate-source voltage V_{gs} and can be expressed as

$$I_d = \frac{W}{L} \mu C_{ox} exp\left(\frac{V_{gs} - V_{th}}{nV_T} \right) \left(1 - exp\frac{-V_{DS}}{V_T} \right), \tag{2.6}$$

where W/L is the device aspect ratio, μ the mobility, C_{ox} the oxide capacitance, V_{th} the threshold voltage, V_{ds} the drain-source voltage, V_T the thermal voltage (kT/q), where q is the electron charge, T the absolute



(a) I_c vs. V_{be} for PNP transistors.

(b) I_{ds} vs. V_{gs} for DTMOS.

Figure 2.3: Cryogenic temperature behaviour of BJT and DTMOS (Images reproduced from [13].)

temperature and k the Boltzmann constant. n is the non-ideality factor given as

$$n = \frac{C_{ox} + C_{depl}}{C_{ox}}, \tag{2.7}$$

where C_{depl} is the depletion capacitance. If V_{ds} larger than a few V_T, rewriting 2.6 yields

$$V_{gs} = V_{th} + nV_T ln\left(\frac{I_D}{\mu C_{ox}(n-1)V_T^2 \frac{W}{L}}\right). \tag{2.8}$$

V_{gs} is empirically showing a CTAT behaviour. The PTAT voltage can be generated, similar as in the case of BJT's, by biasing two MOS transistors in weak inversion with different current densities as shown in figure 2.4. The PTAT voltage can be obtained by taking the difference in their V_{gs},

$$\Delta V_{gs} = V_{gs1} - V_{gs2} = n\frac{kT}{q} ln(p), \tag{2.9}$$

where p is the ratio in bias current between transistor Q_1 and Q_2. The ratio p is a design choice, and can be set by biasing two core devices with a ratio in currents or using the same biasing current but with different core device sizes. By summing this PTAT (2.9) and CTAT (2.8) voltage with a proper scaling factor α, the reference voltage V_{ref} can be obtained,

$$V_{ref} = \alpha \cdot \Delta V_{gs} + V_{gs} = \alpha \cdot n\frac{kT}{q} \ln(p) + V_{gs}. \tag{2.10}$$

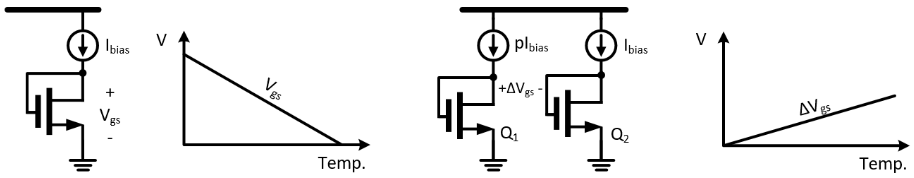


Figure 2.4: Generation of the PTAT- and CTAT voltage using MOS transistors.

Compared with the BJT-based references, MOS-based references are fundamentally extracting the threshold voltage of the transistors (extrapolating at 0 K). However, unlike the bandgap voltage, which is a material constant and hence not depending on the process corner, the threshold voltage highly depends on doping concentration and thus usually has a large spread across corners. As a result, without performing any compensation, the accuracy of the MOS-based references is often worse than the bandgap references [15].

2.2. Key Performance Parameters

The PVT-independence of a voltage reference can be evaluated by a set of key performance metrics. These primary performance metrics include:

- **Temperature coefficient (TC):** The temperature coefficient indicates the drift in the reference voltage over temperature. It is usually calculated using the box method given in equation 2.11, which expresses the maximum difference in V_{ref} over the entire operating temperature range in ppm/K (parts-per-million per kelvin) [16].

$$TC = \frac{V_{ref,max} - V_{ref,min}}{(T_{max} - T_{min}) \times V_{ref,ideal}} \times 10^6. \tag{2.11}$$

- **Inaccuracy (3σ):** The inaccuracy reflects the sensitivity of the voltage reference to process variations. It is usually expressed in 3σ, in which σ is the maximum of the standard deviation based on all the samples over the full temperature range. Note that in order to have sufficient statistical numbers to quantify the inaccuracy, the number of measured samples is also an important consideration for the error bars on 3σ.

- **Noise:** Noise can be seen as a random variation in the output of a voltage reference over time. Noise in the reference voltage is typically compared based on the integrated noise in the 0.1 Hz to 10 Hz as a low-pass filter is usually placed in front of the following circuits to filter out the high-frequency noise.

- **Line regulation:** Line regulation is a measure of the supply voltage dependence of the reference. It is defined as

$$Line\,Regulation = \frac{\Delta V_{out}}{\Delta V_{in}}, \tag{2.12}$$

  where ΔV_{in} is the change in supply voltage and ΔV_{out} is the corresponding change in V_{ref}. In other words, it is the ability of a voltage reference to maintain a constant output voltage despite changes in the supply voltage. It usually has unit V/V or %/V. The smaller this parameter is, the less the reference voltage changes from the nominal value when the supply varies.

As mentioned in the project objective (section 1.4), this project aims to assess the performance limits of cryogenic voltage references. The targeted TC and inaccuracy (3σ) for this project are therefore comparable with the state-of-the-art voltage references operating at room temperatures, which is 10 ppm/K, and 1 %, respectively (as will be introduced in the following section). Noise performance is determined based on the requirements of circuitry in figure 1.1 that will use the reference voltage as input [17, 18]. It is targeted to have 15-bit resolution. Given the expected V_{ref} of 900 mV, the integrated noise in the frequency band from 1 Hz to 10 Hz should remain below 13.7 μV_{rms} to achieve this resolution.

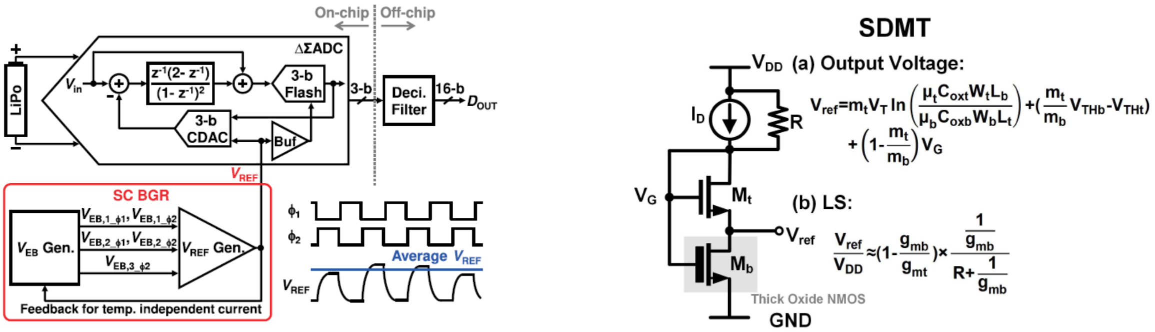
To measure the targeted inaccuracy (3σ) of 1 %, the error caused by supply variations on V_{ref} should be sufficiently small. The allowable error is therefore decided to be 0.1%. For V_{ref}=900 mV, the error on V_{ref} will be smaller than 1mV even if there is a 100 mV change in supply. This translates to 10 mV/V in line regulation. Another parameter that will be taken into account in this project is power consumption. The total available cooling power at 4 K in the dilution refrigerator is only in the order of a few Watts [4]. Power consumption is targeted for 10 µW (0.001 % of the total power budget) in order to allow sufficient power for more power-hungry circuits. The summary of the targeted specifications can be found in table 1.1.

2.3. State-of-The-Art Voltage References - Standard Temperature Range

This section presents an overview of the state-of-the-art voltage references operating on the standard temperature range from −40 °C to 125 °C.

BJT-based voltage references

One of the high-performance voltage references in terms of TC and accuracy is the design proposed by Ge [19]. It was implemented by using a conventional architecture as proposed in [10], but with various compensation techniques such as trimming, chopping, and curvature correction to compensate for process variations and nonlinearities. The up-modulated mismatch errors caused by chopping are filtered by using a switched-capacitor notch filter. Next to this, the voltage reference presented by Boo [20] also employed the above-mentioned compensation techniques. Moreover, this design has dynamic element matching (DEM) applied to the current sources. The architecture of the full system is shown in figure 2.5a. This is a switched-capacitor-based voltage reference, which generates the PTAT- and CTAT voltage in one phase, and sums them together in the next phase. Different DEM arrangements give a slightly different reference voltage, which introduces time-varying ripples. These ripples introduced by DEM are averaged by the oversampling of the sigma-delta ADC that is using the reference. By using these compensation techniques, a TC as low as 5.5 ppm/°C and 3σ of 0.14 % is achieved.


On-chip
Off-chip
ΔΣADC
LiPo
+
−
V_{in}
$\frac{z^{-1}(2-z^{-1})}{(1-z^{-1})^2}$
3-b Flash
3-b CDAC
Buf
3-b
Deci. Filter
16-b
D_{OUT}
V_{REF}
SC BGR
V_{EB} Gen.
$V_{EB,1\_\phi1}, V_{EB,1\_\phi2}$
$V_{EB,2\_\phi1}, V_{EB,2\_\phi2}$
$V_{EB,3\_\phi2}$
V_{REF} Gen.
Feedback for temp. independent current
ϕ_1
ϕ_2
Average V_{REF}
V_{REF}
SDMT
V_{DD}
(a) Output Voltage:
$V_{ref}=m_tV_T\ln\left(\frac{\mu_tC_{oxt}W_tL_b}{\mu_bC_{oxb}W_bL_t}\right)+(\frac{m_t}{m_b}V_{THb}-V_{THt})+\left(1-\frac{m_t}{m_b}\right)V_G$
I_D
R
V_G
M_t
M_b
V_{ref}
(b) LS:
$\frac{V_{ref}}{V_{DD}}\approx(1-\frac{g_{mb}}{g_{mt}})\times\frac{\frac{1}{g_{mb}}}{R+\frac{1}{g_{mb}}}$
Thick Oxide NMOS
GND

(a) Architecture proposed in [20]. (Image reproduced from [20]).

(b) Core architecture proposed in [21]. (Image reproduced from [21]).

Figure 2.5: Architecture proposed in [20] and [21].

MOS-based voltage references

With the trend toward low-power applications, MOS-based references started to draw attention. Shao [21] presents a fully MOS-based voltage reference by stacking the diode-connected MOS transistors. The core architecture is shown in figure 2.5b. It consists of two diode-connected MOS transistors M_t and M_b biased in weak inversion. Given the current flowing through these two transistors is the same, V_{ref} can be expressed as shown in the figure by solving the equation 2.6. By exploring the sizing and the threshold voltage difference of M_t and M_b, the first-order temperature dependence in V_{ref} can be removed. Since the CTAT voltage is generated by exploiting the threshold voltage difference between M_t and M_b, two types of transistors, thick- and thin oxide NMOSs are used to increase the difference. This design achieves a TC of 108 ppm/°C and 3σ of 0.43 % without trimming. It consumes the least amount of power compared with other designs introduced in this section, but TC is also worse. Another voltage reference that was implemented in a comparable technology with this project is presented in [22]. Instead of summing the PTAT- and CTAT voltage together, the transistors are biased in their zero-temperature-coefficient (ZTC) point. At ZTC point, the temperature dependence of carrier mobility and the threshold voltage can compensate for each other. As a result, the voltage that is immune to temperature changes can be obtained by exploring the ZTC point. However, the ZTC point is highly sensitive to process variations, the 3σ of [22] is 1.17 % even after performing 2-point trimming.

The comparison table of the performance for these designs is given in table 2.1. In general, it is possible to achieve high accuracy by using a basic bandgap reference core, but only when multiple compensation techniques are being applied to this core. High-performance designs often come with the cost of high power consumption and complexity. Another thing that is worth noticing is that BJT-based voltage references achieve better performance in terms of TC and 3σ. This is because the bandgap voltage is less prone to the process variations compared with the threshold voltage [21–23].

2.4. Cryo-CMOS and Design Challenges at 4K

State-of-the-art designs already demonstrated high-performance voltage references at room temperature. However, not all of the room temperatures design considerations and approaches are suitable for implementing a wide temperature range voltage reference. Various physical effects cause device behaviour to change when going to cryogenic temperatures. Due to these changes, many proposed architectures for room temperatures applications cannot be used anymore for cryogenic applications. These effects, design considerations, and challenges are briefly presented in this section.

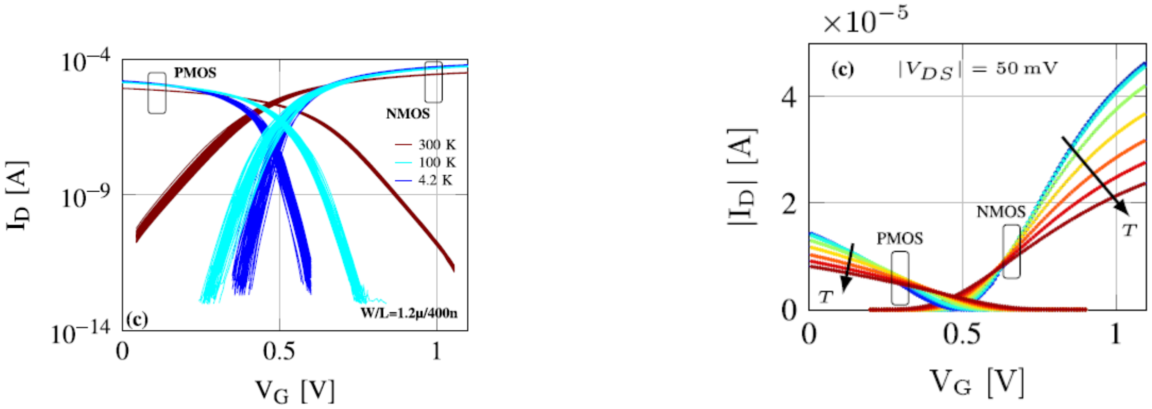
Table 2.1: State-of-the-art voltage references overview for the standard temperature range.

	JSSC 2011 [19]	JSSC 2021 [20]	JSSC 2017 [22]	JSSC 2021 [21]
Technology	0.16um CMOS	0.18um CMOS	65nm CMOS	0.18um CMOS
Supply [V]	1.8	1.8	0.8	1.8
Type	BJT	BJT	MOS	MOS
Temp. range [°C]	-40 to 125	-40 to 125	-40 to 125	-40 to 130
Vref [V]	1.08	1.14	0.69	0.26
TC [ppm/°C]	12	5.5	5.6	108
3σ [%]	0.15	0.14	1.17	0.43
Samples	61	18	50	15
Trimming points	1	1	2	-
Power [W]	99μ	30.6μ	13μ	1.8n
Area [mm^2]	0.12	0.38	0.01	0.006

2.4.1. Device behaviour at cryogenic temperatures

- **Increased threshold voltage:** The threshold voltage is expected to increase 100 mV to 200 mV at 4 K compared to 300 K [24–26]. An increase in threshold voltage implies that a larger V_{gs} is required for the same current. Effectively, this is similar to reducing the supply voltage, hence making low-voltage design at cryogenic temperatures more challenging than at room temperatures.

- **Steeper subthreshold slope:** Figure 2.6a shows the I_d-V_g curves for NMOS device at three different temperature points in the logarithmic scale. The subthreshold slope becomes steeper as temperature goes down [13, 27, 28]. When moving to cryogenic temperatures, the transistors start behaving more as an ideal switch, mainly due to the significantly reduced subthreshold leakage currents. However, for a fixed current, it also causes the required V_{gs} to be higher at cryogenic temperatures than at room temperature when the device is operating in weak inversion. The subthreshold swing (SS) in 40-nm CMOS technology reduces from 90 mV/dec at 300 K to 20 mV/dec at 4.2 K [27].



(a) Id-V_g curves for NMOS device with W/L=1.2 µm/0.4 µm in 40-nm CMOS technology. (Image reproduced from [27]).

(b) Id-V_g curves for N/PMOS devices with W/L=1.2 µm/0.4 µm in 40-nm CMOS technology. (Image reproduced from [24].)

Figure 2.6: DC measurement data of N/PMOS devices in 40-nm CMOS technology.

- **Increased mobility**: Due to the increase in mobility, the drain current I_d at cryogenic temperatures is higher than at room temperatures if the device is operating in strong inversion [29]. For the 40-nm CMOS technology, the mobility increases 2x from 300 K to 4 K [27]. Consequently, the required V_{gs} at 4 K is lower than at 300 K for a fixed current. This is illustrated in figure 2.6b.

- **Increased mismatch:** Mismatch becomes worse at cryogenic temperature [27]. Figure 2.7 shows the NMOS drain current mismatch for different device geometries at three temperature points. The dashed lines in the figure indicate the threshold voltage. Due to the increased threshold voltage, the device is more likely to operate in weak inversion for a given current compared with at room temperatures, which also makes mismatch worse.

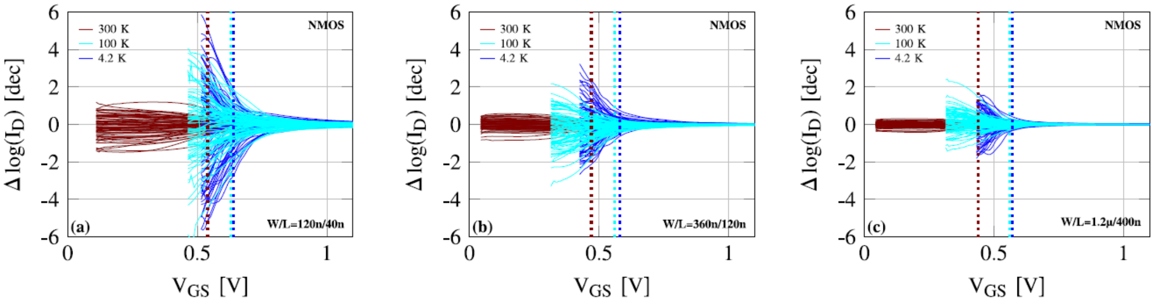


Figure 2.7: NMOS drain current mismatch for different device geometries in 40-nm CMOS technology. (Image reproduced from [27].)

- **Carrier freeze-out:** At cryogenic temperature, dopant does not have enough energy to get ionized. It is referred to as carrier freeze-out, which results in a high substrate resistance [29].

- **Kink effect:** Another effect observed at cryogenic temperatures is the kink effect. When the source-drain voltage is large, the drain current shows a sudden increase. High V_{ds} results in a large electric field, which introduces a bulk current due to the impact ionization. This bulk current, together with the large bulk resistance, results in a significant IR drop that increases the bulk potential which lowers the threshold voltage [29]. Consequently, the drain current increases. However, the kink effect is only observed in mature technologies and therefore is less relevant to this project.

2.4.2. Design considerations and challenges at 4K

The design challenges originating from the devices' behaviour and the general design considerations at 4 K are summarized below.

- **Headroom issue:** An increase in threshold voltage leads to a headroom issue. Consequently, only the architecture that is suitable for low-voltage applications will be considered.

- **Miamatch degradation:** As will be discussed in chapter 3, mismatch degrades the 3σ and TC of the voltage references directly. Therefore, extra care, such as careful sizing/layout or compensation techniques, should be taken to minimize the effects of mismatch.

- **Moving bias points:** Especially in voltage-mode voltage references, where the currents are typically PTAT, the current levels can easily change up to 5x [6] when moving from 300 K down to 4 K. Depending on the exact sizing, transistors may observe major shifts in bias points or even operating regions. This may potentially introduce nonlinear errors and therefore affect the TC.

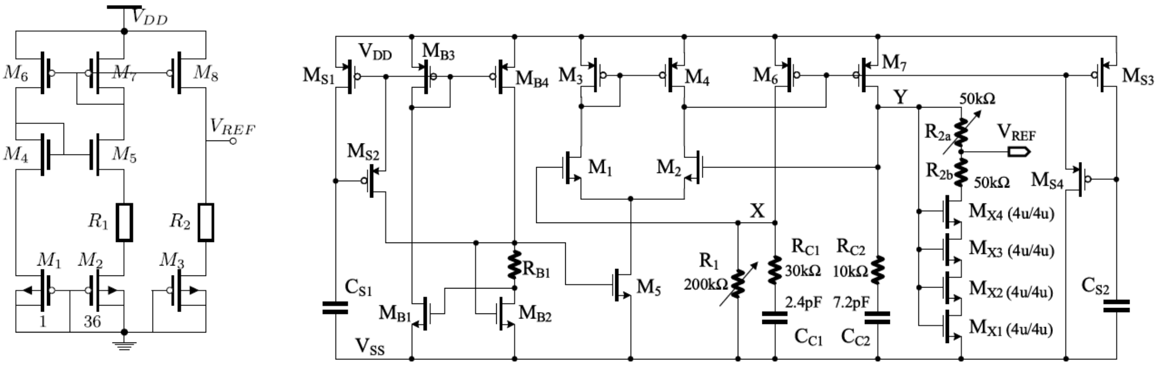
- **Limited power dissipation:** The available cooling power is limited at low-temperatures [4, 30]. As a result, the power budget is less. Using low power to achieve high performance is always a challenge.

- **No reliable models:** High-performance voltage references require highly accurate models for design. Especially when performing the curvature correction and temperature compensation [16], precise models are essential to take all the second-order effects into account. Yet, reliable models at cryogenic temperatures are not available during the design phase of this project, design has to rely on the measurement data. Consequently, extra care has to be taken to ensure the performance at 4 K.

2.5. State-of-The-Art Voltage References - Cryogenic Temperature Range

This section presents an overview of the state-of-the-art voltage references that can operate down to the cryogenic temperature range. Among them, Homulle presented the first MOS-based voltage reference working at 4 K [14]. The design is shown in figure 2.8a. It was realized by using dynamic-threshold MOS (DTMOS) with the approach introduced in section 2.1. In contrast to the commonly used approach, it suggests that biasing transistors in strong inversion can give better performance. The work presented in [6] explores different devices: NMOS, PMOS, and DTMOS, to assess the best candidate for implementing a cryogenic voltage reference. It provides sufficient statistical characterization to give insight into the effects of process variations on cryogenic voltage references. This work will be discussed more in chapter 3. The design proposed by

Yang [31] is based on the ZTC point approach. The architecture is shown in figure 2.8b, where M_X is the core transistor. Opamp is used to equal the voltage at node X and node Y, consequently, the desired bias point of M_X can be reached by choosing the value of R_1. Through the help of R_2, the drain voltage of M_X can also be set properly. M_{Nx} is implemented by stacking identical devices to increase the effective length to reduce the sensitivity of the device to process variations. However, ZTC point is very sensitive to the biasing condition and process variation [22], given the extracted device models at 4 K and 77 K, this design only achieves a TC of 1214 ppm/K from 4 K to 300 K. Moreover, the authors only report measurements for 1 sample, leaving the inaccuracy undefined. Next to this, Liu presents a cryogenic bandgap reference using BJT devices [32]. However, this design only be measured down to 77 K. And since BJTs will lose their characteristic at cryogenic temperature as mentioned in section 2.1.1. Therefore, it is less relevant to this project.



(a) Voltage reference proposed in [14]. (Image reproduced from [14].)

(b) Voltage reference proposed in [31]. (Image reproduced from [31])

Figure 2.8: Architecture proposed in [14] and [31].

A performance comparison is summarized in table 2.2. Among the designs that can work down to 4 K, [6] achieves the best performance in terms of TC and 3σ. Nevertheless, it still cannot reach comparable performance to the designs at room temperatures. This can be attributed to the following two reasons. Firstly, the designs presented so far are mainly based on the basic core devices (as shown in figure 2.4) and without using further compensation techniques. As a result, the intrinsic accuracy and temperature dependence of the devices will limit the performance directly. Secondly, there are no reliable models to predict what happens at cryogenic temperatures, which adds another challenge when aiming for high performance.

Table 2.2: State-of-the-art voltage references overview for cryogenic temperature range.

		SSCL 2018 [14]	ESSCIRC 2019 [6]			SSCL 2020 [31]	TNS 2020 [32]
Technology		40nm CMOS	40nm CMOS			28nm FDSOI	0.18um CMOS
Supply [V]		3.3	1.1			1.2	1.8
Type		DTMOS	NMOS	PMOS	DTMOS	MOS	BJT
Temp. range [K]		4 to 320	66 to 300	4 to 300		4 to 300	77 to 290
Vref [V]	T_{max}	1.02	0.48	0.54	0.63	0.49	1.17
	T_{min}	0.81	0.48	0.71	0.6	0.66	1.14
TC [ppm/K]		833	76	539	436	1214	102.8
3σ [%]		6.6	1.2	2.2	1.7	NA	1.29
Samples		NA	14	28	28	NA	5
Trimming points		NA	1			NA	-
Power [W]	T_{max}	368μ	12.3μ	13.8μ	13.9μ	15.8μ	3.24μ
	T_{min}	132μ	6.5μ	7.4μ	7.3μ	13.9μ	0.72μ
Area $[mm^2]$		0.0004	0.006	0.009	0.009	0.041	NA

As mentioned above, one of the missing parts in the state-of-the-art of cryogenic voltage references is the implementation of the compensation techniques. It would potentially be a bottleneck that limits the performance of the cryogenic voltage references. To fill this gap and to know what kinds of errors should be compensated for, chapter 3 will introduce different error sources present in the MOS-based voltage reference. Furthermore, the effect of different compensation techniques on performance will be investigated.

3

Intrinsic Accuracy of the Voltage Reference

3.1. Introduction

As mentioned in chapter 2, one of the design challenges for cryogenic electronics is to cope with the lack of CAD-compatible cryogenic device models. To point out the direction for system- and circuit design, the main objective of this chapter is to identify the bottlenecks for designing cryogenic voltage references, and assess the performance that can be achieved when applying certain compensation techniques. This analysis is based on datasets from earlier voltage reference designs. The organization of the chapter is as follows. Firstly, the available datasets will be introduced on which the analysis is based. Secondly, the temperature dependence of the PTAT- and CTAT voltage will be explored. Finally, the effect of process variations on V_{ref} will be introduced, followed by the compensation techniques that allow for mitigating the error sources.

3.2. Available Datasets

The architecture of the reference generator used in this project is the same as the voltage reference presented in [6], as it has already been proven to work from room temperature down to 4 K. The simplified architecture is shown in figure 3.1. The schematic on the left is the reference with NMOS as core devices, while the right is the reference with PMOS as core devices. This circuit consists of a PTAT generator formed by M_1, M_2 and R_1. With a current density ratio p in the core transistors M_1 and M_2, which is set by the current sources M_3 and M_4, a PTAT voltage is generated across R_1. This PTAT voltage is then been converted into current and mirrored to the output branch through M_5, which also amplifies the current with a factor m. The PTAT voltage is now effectively scaled-up by a factor $(\gamma)= m \cdot R_2/R_1$. The reference voltage can be obtained by adding this scaled PTAT voltage to the CTAT voltage generated by M_6. More details will be presented in chapter 4. In the following contents, R_1 will be referred to as the PTAT resistor R_{ptat} and R_2 will be referred to as the output resistor R_{out}.

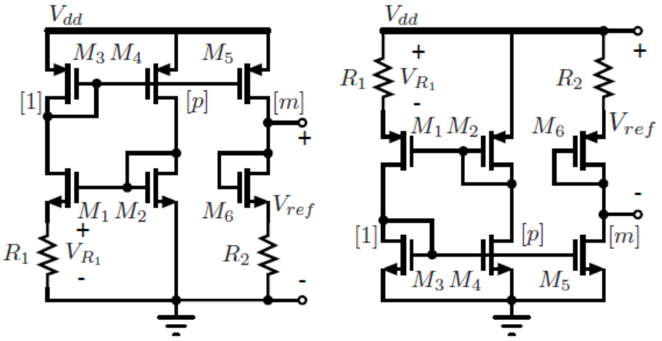
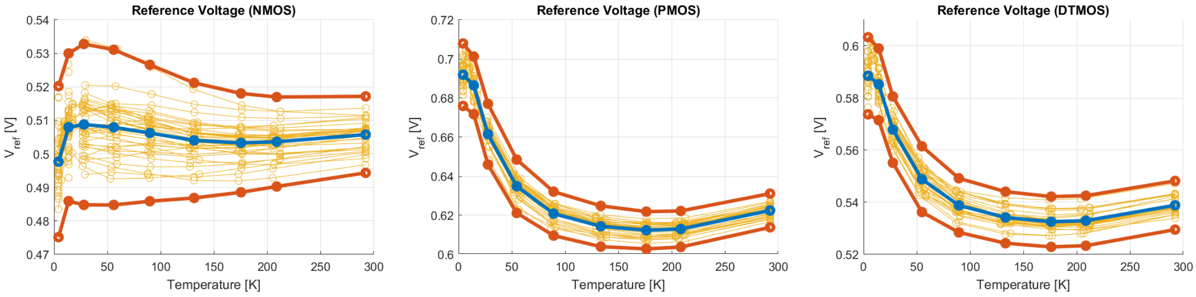


Figure 3.1: Simplified architecture of the voltage reference proposed in [6]. (Image reproduced from [6].)

The measurements presented in this section originate from an extension of the work presented in [6]. These measurements were already available at the start of this project. The circuits that were measured were from two batches, where each batch has three flavors of voltage references on-chip, namely NMOS-, PMOS-, and DTMOS-based references. Different flavors of voltage references have the same architecture as shown in figure 3.1, the difference only lies in the types of the core devices ($M_{1,2,6}$). The measured reference voltage is shown in figure 3.2. Yellow curves indicate the measurement data, while the blue curve is the average reference voltage and the red curves the $\pm 3\sigma$. The temperature coefficients of these three types of references are 347 ppm/K, 559 ppm/K, and 484.2 ppm/K, respectively, calculated by the box method as shown in equation 2.11. The inaccuracy (3σ) is 4.8 %, 2.6 %, and 3 %, respectively. In order to characterize and investigate the error sources contributing to the inaccuracy, an architecture that allows for DEM on both the current sources and core transistors has also been taped out and measured. There are 36 samples for each flavor of the voltage reference and 16 samples for the architecture that includes the switches for DEM.



(a) NMOS-based reference.

(b) PMOS-based reference.

(c) DTMOS-based reference.

Figure 3.2: Measured reference voltage from 36 samples for the NMOS-, PMOS-, and DTMOS-based references.

3.3. Temperature Dependence of the PTAT- and CTAT Voltage

The working principle of the voltage reference presented in [6] is based on combining a PTAT voltage with a CTAT voltage. Therefore, a low temperature-drift reference voltage can only be generated if the PTAT- and CTAT voltage are sufficiently linear. However, the measurement data shows there is a large curvature present at cryogenic temperatures. To investigate the origin of this nonlinearity, the PTAT- and CTAT voltages are examined separately. In this section, only the effect of nonlinearity will be analyzed, i.e., the effect of mismatch has not been taken into account.

3.3.1. PTAT from the measurements

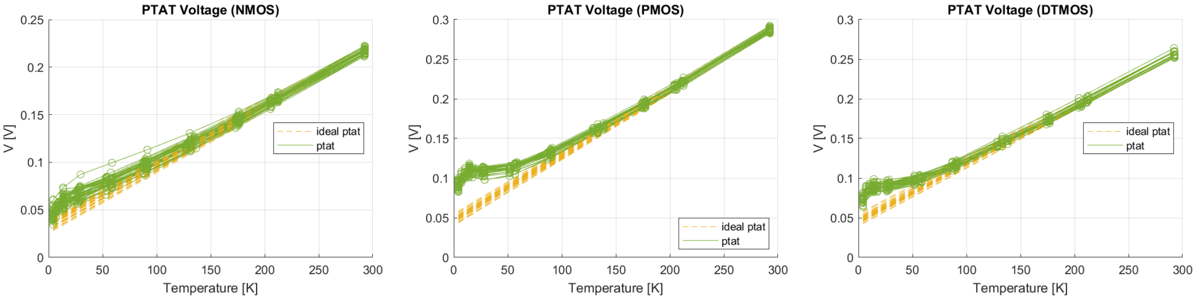
The ideal PTAT voltage V_{ptat} can be expressed as

$$V_{ptat} = n\frac{KT}{q}ln(p), \tag{3.1}$$

where p is the current density ratio between M_1 and M_2 in figure 3.1 and n the subthreshold non-ideality factor. The measured V_{ptat} is shown in figure 3.3 below. The green lines represent the measured data points while the yellow lines indicate the ideal PTAT trend, obtained by extrapolating the data between 220 K and 300 K. In all the measured curves, for all three device flavours, V_{ptat} shows a similar trend. Starting from 300 K, V_{ptat} first decreases linearly with temperature as predicted by equation 3.1, and starts to saturate when reaching approximately 60 K. Saturation remains until around 15 K and a kink-like behaviour appears. Note that the kink here refers to the observation in the measurements of the PTAT voltage in figure 3.3, which is different from the so-called kink effect at cryogenic temperatures [29].

As mentioned in chapter 2, the generation of V_{ptat} is based on the operation of the core transistors in the subthreshold region. The drain current can then be approximated as in equation 2.6. The subthreshold swing (SS), which is the required change in gate voltage for 10× change in current can be computed as

$$SS = \left(\frac{\partial I_D}{\partial V_G}\right)^{-1} = n\frac{kT}{q}ln(10), \tag{3.2}$$



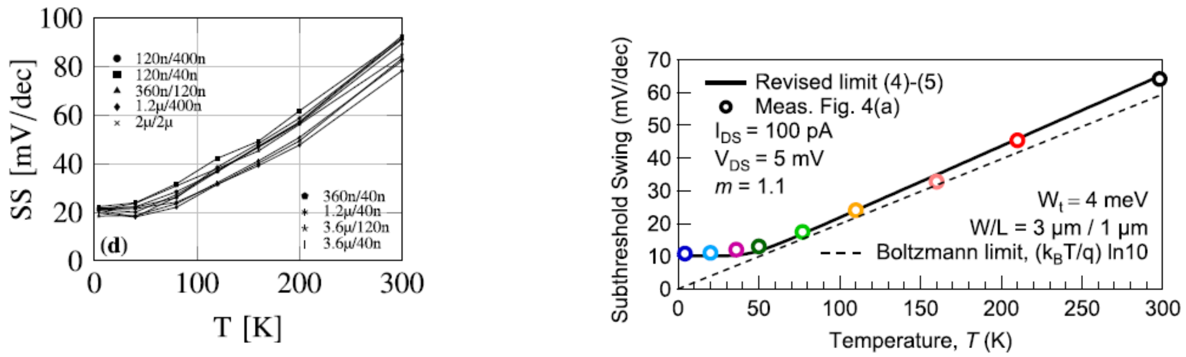
(a) V_{ptat} from NMOS-based reference.

(b) V_{ptat} from PMOS-based reference.

(c) V_{ptat} from DTMOS-based reference.

Figure 3.3: Measured PTAT Voltage for the NMOS-, PMOS-, and DTMOS-based references.

usually expressed in mV/dec. The SS determines the temperature dependence of V_{ptat}. In other words, V_{ptat} follows the trend of the subthreshold slope. Figure 3.4 shows the extracted subthreshold swing in 40-nm and 28-nm CMOS technology as found in literature. The saturation of the subthreshold swing is observed, essentially regardless of the technology [13, 14, 27, 33]. Given that k, q in equation 3.1 are physical constants, the saturation of V_{ptat} below 60 K might be caused by the increase in n, provided that the equation 3.2 still holds at cryogenic temperatures. In case equation 3.2 still holds at cryogenic temperatures, n increases from approximately 1.7 at room temperature to larger than 50 at 4 K, indicating that C_{dep} significantly increases at cryogenic temperatures, which is not reasonable and has been proved wrong [28]. Another explanation for this might be associated with the imperfection of the interface or the blurring in the band-tail [28] that makes the subthreshold swing differ from the prediction.



(a) Measured subthreshold swing in 40-nm CMOS technology. (Image reproduced from [27].

(b) Measured subthreshold swing in 28-nm CMOS technology. (Image reproduced from [28].)

Figure 3.4: Measured non-ideality factor and subthreshold swing vs. temperature.

Another nonlinearity that cannot be described by equation 3.1 is the kink around 15 K. This kink is observed in both batches of the references, yet it is still not clear what is the actual cause. Figure 3.5 shows the measured V_{ptat} from another PMOS-based voltage reference design. The architecture of this voltage reference is the same as the others, but the absolute length of the core device was reduced from 2 μm to 0.4 μm. There are in total 4 samples for this reference design. Compared with figure 3.3, the kink has been alleviated significantly by using this shorter device and even disappears in one of the samples. Kink might be caused by circuit-related issues or the physical behaviour that is associated with the channel length. However, further measurements are required in order to draw a conclusion on what is the exact cause.

3.3.2. CTAT from the measurements

The measured CTAT voltage is shown in figure 3.6. The green lines show the measurement results while the dashed yellow lines represent the ideal CTAT voltage, interpolating from the measurements at room temperatures. The general trend is that V_{ctat} is linear with the temperature down to around 70 K, from where it starts to deviate from the linear trend. V_{ctat} of the NMOS-based references saturates, while V_{ctat} of the P/DTMOS-based references shows a sudden increase. V_{ctat} used in this design was taken from the V_{gs} of the output

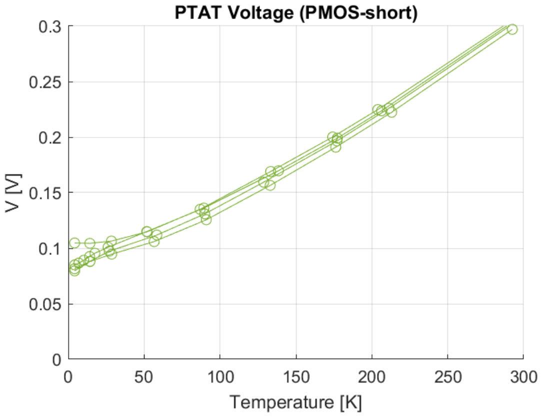
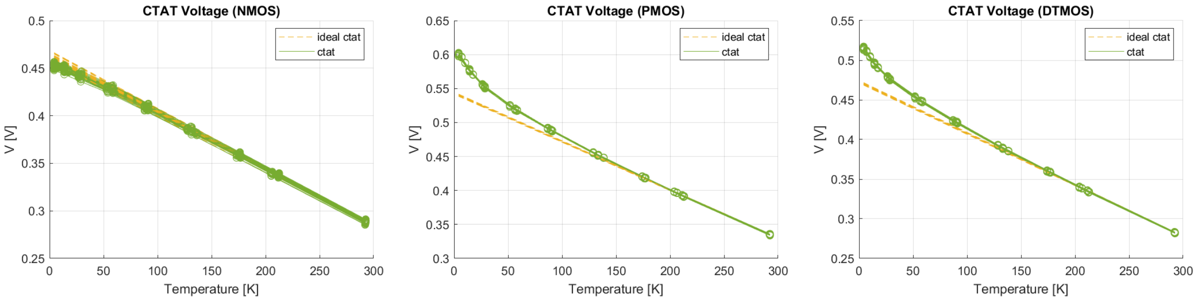


Figure 3.5: Measured PTAT voltage of the short channel PMOS-based reference.

transistor. V_{gs} mainly follows V_{th} as can be seen from equation 2.8. Therefore, the nonlinear trend in V_{ctat} is mostly caused by the non-linear behaviour in V_{th}.



(a) V_{ctat} from NMOS-based reference.

(b) V_{ctat} from PMOS-based reference.

(c) V_{ctat} from DTMOS-based reference.

Figure 3.6: Measured CTAT Voltage for the NMOS-, PMOS-, and DTMOS-based references.

The trend of V_{ctat} from the measurements is similar to the trend of V_{th} found in literature. Figure 3.7 shows measurement data of the threshold voltage in TSMC's 40-nm and 28-nm technologies. For the NMOS devices, and PMOS devices with smaller sizes, V_{th} saturates when the temperature is below 50 K. V_{th} can be described as

$$V_{th} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}}, \tag{3.3}$$

where Φ_F is defined as

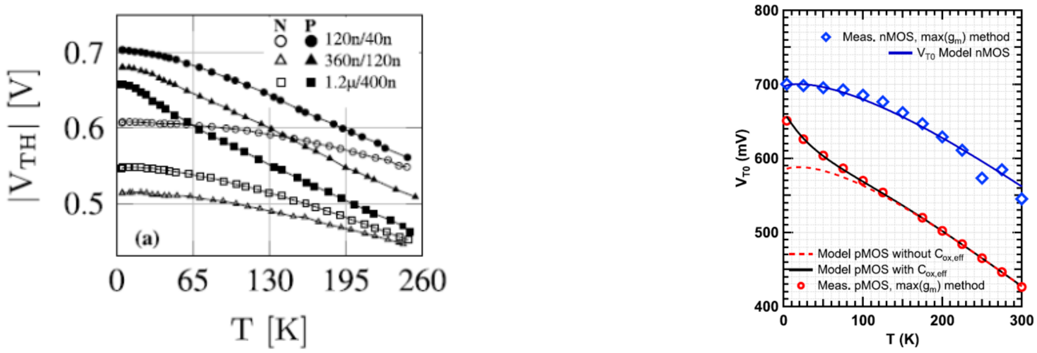
$$\Phi_F = \frac{kT}{q} ln\left(\frac{N_{sub}}{n_i}\right), \tag{3.4}$$

and where $\Phi_{MS} = \Phi_M - \Phi_S$ is the work function difference between the gate (Φ_M) and the semiconductor (Φ_S), N_{sub} the doping concentration of the substrate, n_i the intrinsic carrier concentration, C_{ox} the oxide capacitance per unit gate area and Q_{dep} the charge in the depletion region [34]. To first order, V_{th} has a CTAT behaviour. That is because the Fermi level is closer to the valence band when going down in temperature due to the scaling of the Fermi-Dirac occupation function. As a result, a higher gate voltage is required for band bending to push the Fermi level closer to the conduction band [25, 26]. n_i in equation 3.4 can be described as

$$n_i = \sqrt{N_c N_v} exp\left[\frac{-E_g(T)}{2kT/q}\right], \tag{3.5}$$

where N_c is the density-of-states in the conduction band and N_v is the density-of-states in the valence band. The exponential term inside n_i makes V_{th} saturate after reaching a certain temperature. Yet, the threshold voltage of PMOS devices exhibits a sudden increase below 70 K instead of saturation. This might be attributed to the temperature dependence of the effective C_{ox}, as it decreases exponentially below a certain temperature [26]. The dashed and the solid lines in figure 3.7b show the model for V_{th} with and without including the effective value of C_{ox}. By including C_{ox}, V_{th} of the PMOS devices can be well-predicted [26]. However, it is not yet understood why the effective C_{ox} of PMOS suddenly drops below a critical temperature.

In addition to the temperature dependence of the threshold voltage itself, there are higher-order temperature-dependent terms present in V_{gs} and have an effect on V_{ctat} as indicated in equation 2.8. Such an error can be compensated for by using curvature correction techniques. For example, an additional nonlinear term can be generated with an opposite sign to compensate for the nonlinearity in V_{gs} [16]. Besides, since the biasing current in the reference I_{ptat} is not linear it will make V_{ctat} slightly off from the desired value. However, these higher-order errors are still negligible compared with the nonlinearity in V_{th} at cryogenic temperatures.



(a) Threshold voltage of N/PMOS devices with different geometries in 40-nm CMOS technology. (Image reproduced from [24].)

(b) Threshold voltage of N/PMOS devices with W/L=10 μm/1 μm in 28-nm CMOS technology. (Image reproduced from [26].)

Figure 3.7: Measured threshold voltage in 40-nm and 28-nm CMOS technology.

3.3.3. Summary

For obtaining a reference voltage with low temperature drift, it is required that the PTAT- and CTAT voltage are sufficiently linear. However, both in state-of-the-art cryogenic voltage references [6] and in the extended measurement data, significant nonlinearities are present. The previous section presented the measured PTAT- and CTAT voltage and investigated their nonlinearities. Some observations and guidelines are given below.

- PTAT voltage: In general, the PTAT voltage saturates below 50 K due to the saturation of subthreshold swing. Furthermore, it exhibits a kink-like feature around 15 K. The saturation of the subthreshold swing is unavoidable, while the kink can be alleviated by using a shorter channel length of 0.4 μm instead of 2 μm.

- CTAT voltage: The temperature dependence of V_{ctat} is dominated by V_{th}. The CTAT voltage from the NMOS-based references shows a different trend from DT/PMOS-based references. For the NMOS, the CTAT voltage starts saturating below 50 K while PMOS and DTMOS show an increase in CTAT voltage below 50 K.

- Due to physical limitations, it is not possible to generate a PTAT- and CTAT voltage that is sufficiently linear for the temperature range from 4 K to 300 K. However, if V_{ptat} and V_{ctat} both saturate at cryogenic temperatures, the nonlinearity in the PTAT- and CTAT voltage can cancel out each other.

Based on the linearity of the PTAT- and CTAT voltage, NMOS transistors will be used as the core device in this design. To further investigate whether the channel length is the cause of the kink in the PTAT voltage, two versions of the reference generator will be implemented, one with W/L of 15 μm/0.4 μm (v1) and the other with W/L of 15 μm/2 μm (v2).

3.4. Effects of Compensation Techniques on TC and Variation

Due to process variations, the measured reference voltage for each measured reference voltage has a statistical deviation from the nominal value. This can be seen from the discrepancies between samples in the reference voltage in figure 3.2. Some major error sources in the typical CMOS voltage references include threshold voltage mismatch/spread, current source mismatch, and resistor mismatch/spread [16, 19, 23]. Fortunately, there are several techniques, such as chopping, dynamic element matching (DEM), and trimming, that can be used to alleviate the effects of process spread and mismatch.

Reducing the mismatch between samples does not only help to improve the accuracy but in some cases also lowers the TC, since some of the error sources are temperature-dependent. To even further improve the performance, curvature correction might be required to cancel the higher-order nonlinearity present in the CTAT voltage [19]. However, applying such techniques often comes at the cost of increased power consumption and circuit complexity. This section aims at gaining an understanding of how the error sources affect the reference voltage, and up to what extent the compensation techniques improve performance. Firstly, the dominant error sources are examined based on the measurement data and theory. Secondly, different compensation techniques will be investigated, from which their effectiveness is assessed based on the improvement in terms of TC and variation.

3.4.1. Error sources

This sub-section introduces the main error sources for the architecture in figure 3.1, including current source mismatch, core transistor mismatch, resistor mismatch, output transistor spread, and resistor spread. Although these error sources exhibit different temperature dependencies, they can generally be categorized into three types: PTAT errors, offset errors, and non-linear errors. The error sources are translated to V_{ref} through the sensitivity functions derived in appendix A.

Current source mismatch

Accurate generation of the reference voltage requires accurate current ratios for biasing and mirroring. Current sources usually operate in strong inversion for better current matching [35, 36]. The drain current of the MOS transistor when it is in strong inversion can be approximated as

$$I_d = \frac{\beta}{2}(V_{gs} - V_{th})^2(1 + \lambda V_{ds}), \tag{3.6}$$

where β is given as $\mu C_{ox} W/L$ and λ is the channel-length modulation factor. Due to the threshold voltage mismatch, β-mismatch, and finite output impedance of the current sources, the actual current ratio often differs from the desired one. In the architecture proposed in [6], there are mainly two current ratios involved, p and m, with respect to the unit-size current source. That is, $p = I_{d4}/I_{d3}$ and $m = I_{d5}/I_{d3}$. If we define $p' = p + \delta p$, then ΔV_{gs} becomes

$$\Delta V'_{gs} = n\frac{kT}{q} ln(p + \delta p) = n\frac{kT}{q} ln\left[p\left(1 + \frac{\delta p}{p}\right)\right] \approx n\frac{kT}{q}\left[ln(p) + \left(\frac{\delta p}{p}\right)\right], \tag{3.7}$$

where δp is the error in the current ratio p. Compared with the ideal ΔV_{gs}, when there is no mismatch in the current sources, the error in ΔV_{gs} caused by δp can be derived as

$$\delta\Delta V_{gs} = \Delta V'_{gs} - \Delta V_{gs} = n\frac{kT}{q}\frac{\delta p}{p}. \tag{3.8}$$

This error will be translated to V_{ref}, causing an error δV_{ref} through the sensitivity function A.4,

$$\delta V_{ref} = n\frac{kT}{q}\frac{\delta p}{p} \times S^{V_{ref}}_{\Delta V_{gs}} = n\frac{kT}{q}\frac{\delta p}{p}\left(\gamma + \frac{1}{ln(p)}\right), \tag{3.9}$$

where γ is the scaling factor

$$\gamma = m \cdot \frac{R_{out}}{R_{ptat}}. \tag{3.10}$$

The current ratio m is mainly used to mirror the PTAT current to the output branch, where it is used as part of the scaling factor (γ). Consequently, an error in m causes an error in γ. Assuming the error ratio in m can be modeled as $m' = m + \delta m$, the scaling factor becomes

$$\gamma' = (m + \delta m) \times \frac{R_{out}}{R_{ptat}} = m\left(1 + \frac{\delta m}{m}\right)\frac{R_{out}}{R_{ptat}}. \tag{3.11}$$

If the error in the scaling factor is modeled as $\delta\gamma = \gamma\delta m/m$, its effect on V_{ref} can be derived as follows,

$$\delta V_{ref} = \gamma\frac{\delta m}{m} \times S_{\gamma}^{V_{ref}} = \gamma\frac{\delta m}{m} \times n\frac{kT}{q}\left(ln(p) + \frac{1}{\gamma}\right). \tag{3.12}$$

Based on room temperature simulations, current source mismatch has the largest effect on V_{ref} within a single corner with respect to other error sources introduced in section 3.4.1. As mentioned in section 3.2, the design presented in [6] features DEM on the current sources, which allows rearranging the order of the current sources. With the current ratio in M_3:M_4:M_5 being 1:p:m=1:10:5, there are in total 16 current sources in the architecture presented in figure 3.1. The V_{ref} of different arrangements of the current sources at 4 K is shown in figure 3.8. The x-axis indicates different arrangements. For a given chopping phase, the voltage difference between each phase represents the effect of the current source mismatch. More details regarding the implementation of DEM and chopping will be introduced in section 3.4.2.

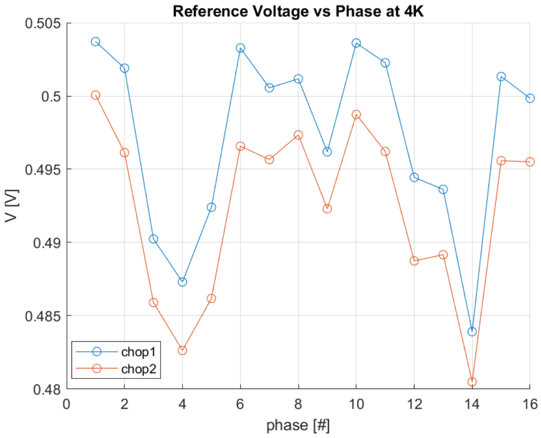
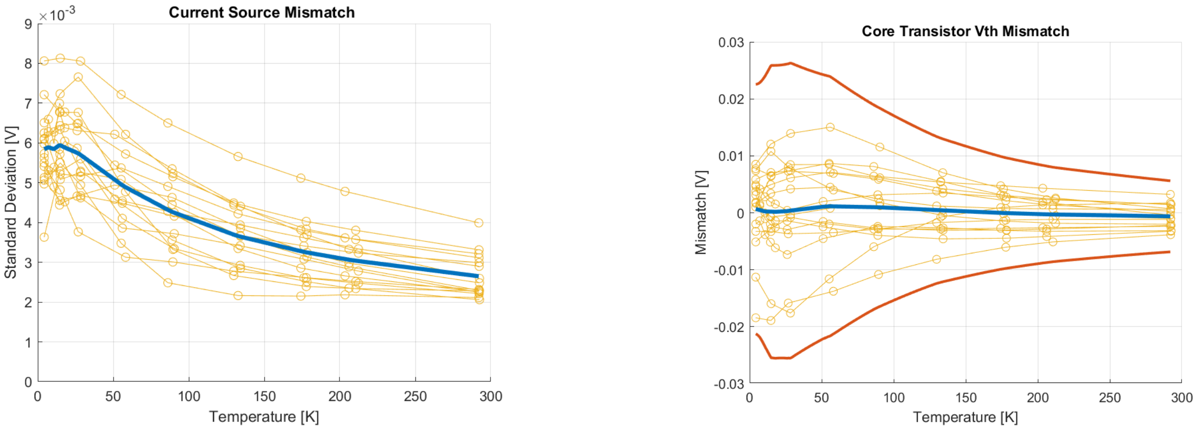


Figure 3.8: Reference voltage for all measured DEM/chopping arrangements at 4 K.

Figure 3.9a shows the extracted variation associated with the current sources. It was derived from figure 3.8 by computing the standard deviation of the 16 arrangements. This procedure was repeated for 16 samples at all the measured temperature points. The current source mismatch increases as the temperature goes down. In the voltage reference architecture, the biasing current has a PTAT nature. For such a wide temperature range, a 5x current reduction is observed in the measurements. Current sources are therefore expected to enter weak inversion at cryogenic temperatures due to the reduced current. As a result, mismatch increase significantly [27]. Although equation 3.9 and equation 3.12 seems to suggest a PTAT-error if δp and δm are constant, this is not the case due to the change in the operation region. Such an error therefore causes additional temperature drift in V_{ref}.

Threshold voltage mismatch/spread

Assuming there is no mismatch, the ideal PTAT voltage ΔV_{gs} can be expressed as in equation 2.9. However, since the bulk of M_1 and M_2 is grounded, but the potential at the source of M_1 is defined by ΔV_{gs}, the body effect will introduce a systematic threshold voltage mismatch between them. In addition, process variations will also cause a random mismatch. The mismatch $\delta V_{th} = V_{th,M1} - V_{th,M2}$ in the PTAT voltage will be converted into an error in bias current I_{ptat}. Therefore, it not only causes the error in PTAT voltage but also alters the biasing condition of the output transistor M_6, introducing a shift in CTAT voltage. This error will be transferred to the output through the sensitivity function A.4, generating an error in V_{ref} which can be approximated as



(a) Current source mismatch.

(b) Threshold voltage mismatch in core devices.

Figure 3.9: Effective mismatch in core transistors and current sources.

$$\delta V_{ref} = \delta \Delta V_{gs} \times S_{\Delta V_{gs}}^{V_{ref}} \approx \delta V_{th} \times \left(\gamma + \frac{1}{ln(p)}\right). \tag{3.13}$$

Figure 3.9b shows the threshold voltage mismatch between the core devices versus the temperature, extracted from the measurements. The data was extracted by taking the difference in V_{ref} from the two chopping phases. Given that ΔV_{gs} will be amplified by the scaling factor γ, this extracted quantity is therefore not δV_{th} between $M_{1,2}$ but the scaled-up of δV_{th}. In the general trend, the threshold voltage mismatch increases as the temperature decreases. Note that all the mismatch in the core transistors was combined into an equivalent threshold voltage mismatch in this extraction approach. β-mismatch of the core transistors is thus also included in it. Given the sizing that was used in [6], γ=3.1 and p=10, δV_{th} will be amplified approximately 3.5 times, thus degrading the TC on a significant scale.

Another non-negligible error source in the voltage reference is the threshold voltage spread of the output transistor M_6. The reference voltage generated by MOS-based references is directly related to the threshold voltage of the transistor. The process sensitivity of the output transistor is thus very important [21, 23]. The threshold voltage spread in $M_{1,2}$ will not introduce errors as the spread in $M_{1,2}$ will cancel out each other. However, this is not the case for the output transistor. As derived in equation A.3, the spread in the output transistor will be added to the reference voltage directly,

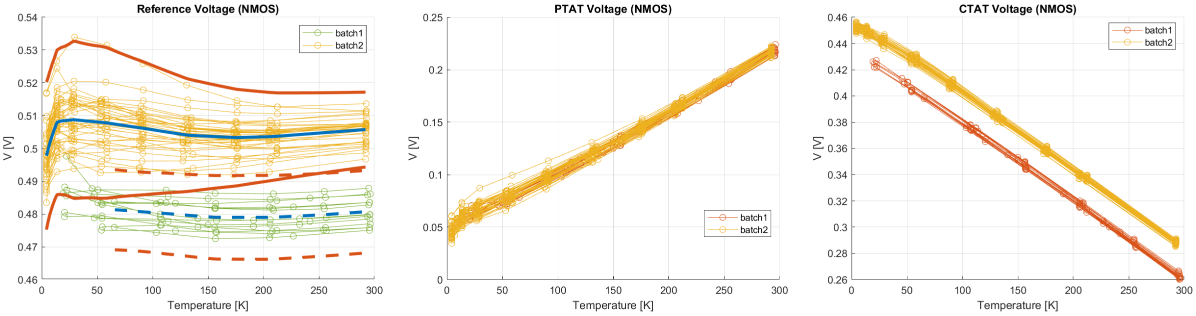
$$\delta V_{ref} = \delta V_{th6} \times S_{V_{gs}}^{V_{ref}} \approx \delta V_{th6}. \tag{3.14}$$

For a device size of W/L=120 µm/0.4 µm in TSMC 40-nm CMOS technology, the threshold voltage spread is around 4 mV in a single corner. Nevertheless, the spread can easily be in the order of 35 mV across the corners. Figure 3.10 shows the reference, PTAT-, and CTAT voltage from two batches. The PTAT voltage remains the same for two batches while the CTAT voltage has a 30 mV discrepancy. The observation is in agreement with the expectation that the PTAT voltage is less pronounced to the process spread as mentioned above. The final goal is to design a batch-independent voltage reference, hence the threshold voltage spread is a non-negligible error source that has to be compensated for.

Resistor mismatch/spread

Resistor mismatch refers to the mismatch between R_{ptat} and R_{out}. Since this ratio determines the scaling factor, first-order temperature dependence in PTAT and CTAT voltage cannot be compensated for each other if the resistor ratio is not accurate. Let $\alpha = R_{out}/R_{ptat}$ be the ratio between R_{ptat} and R_{out}, and $\Delta\alpha$ the error caused by resistor mismatch. The scaling factor becomes $\gamma' = m(\alpha + \delta\alpha)$. It generates an error voltage in V_{ref} through sensitivity function A.5

$$\delta V_{ref} = \delta\gamma \times S_{\gamma}^{V_{ref}} = \gamma\frac{\Delta\alpha}{\alpha} \times S_{\gamma}^{V_{ref}} = \gamma\frac{\Delta\alpha}{\alpha} \times \left(n\frac{kT}{q}\left(ln(p) + \frac{1}{\gamma}\right)\right). \tag{3.15}$$



(a) Reference voltage.

(b) PTAT voltage.

(c) CTAT voltage.

Figure 3.10: Batch comparison of reference-, PTAT-, and CTAT voltage.

As the first order temperature dependence in the two resistors cancel out each other in the ratio, $\Delta\alpha$ is up to first order temperature independent. Mismatch in resistors will therefore translate to a PTAT error in V_{ref}. However, the change in the absolute value of R_{ptat} will alter the biasing current of the output transistors M_6, causing a shift in V_{gs6}. Fortunately, since the biasing current shows a PTAT behaviour, this error can be seen as a PTAT error and can therefore be trimmed out.

3.4.2. DEM and chopping

This section covers two dynamic offset-cancellation techniques: chopping and DEM. The working principle of them will be introduced first, followed by their effect on TC and variation.

Dynamic element matching (DEM)

DEM is a technique that dynamically interchanges identical circuit elements, which effectively up-converts the mismatch error. By doing so, the mismatch errors can be averaged out over time which significantly reduces the error [16, 20]. To generate the accurate biasing current ratio, DEM can be applied to the current sources M_{3-5} in figure 3.1. The current ratio for M_3, M_4 and M_5 is 1:10:5. According to the Pelgrom law, the unit size current source in the left-most branch is the main error contributor as it has the smallest absolute size [36]. Therefore, the idea is to average the mismatch error by swapping the position of the current sources in the circuit. Each current is then used as the unit-current source in the left-most branch once. By dynamically arranging the elements, the mismatch is up-converted to the harmonics of the DEM frequency and can be filtered out by using a low-pass filter.

Chopping

The principle of chopping is to use the modulation technique to separate the frequency of the signal and the offset, and then use the filtering method to remove the offset while the signal is unaffected [16]. The principle is illustrated in figure 3.11. The desired signal is up-modulated by the first chopper to the chopping frequency before entering the amplifier. In this way, the desired signal and error signal (V_{os}) will be in different frequency bands. After the amplifier, the second chopper de-modulates the input signal back to the baseband and up-modulates the offset to the chopping frequency. Similar to DEM, the up-modulated offset can be filtered out by the low pass filter after the second chopper.

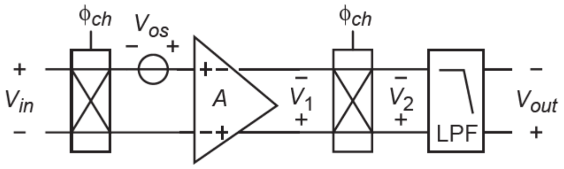


Figure 3.11: Typical implementation of chopping. (Image reproduced from [16].)

Chopping is usually applied to the input of the amplifier to reduce the offset and the low-frequency noise from the input pair. This concept can be applied to the core device $M_{1,2}$ in figure 3.1 as well. The basic idea here is to dynamically interchange the core devices. By doing so, the mismatch between $M_{1,2}$ (δV_{th}) is up-modulated to the chopping frequency. However, note that in this design, the desired signal is always at the baseband and thus the second chopper is not required. Therefore, strictly speaking, the chopping used here is essentially applying DEM on the core devices.

The effect of chopping can also be seen from another perspective. Using the same assumption above, if M_1 has a δV_{th} mismatch with respect to M_2, then during chopping phase 1, the PTAT voltage can be expressed as

$$V_{ptat,chop1} = \delta V_{th} + n\frac{kT}{q} ln(p), \tag{3.16}$$

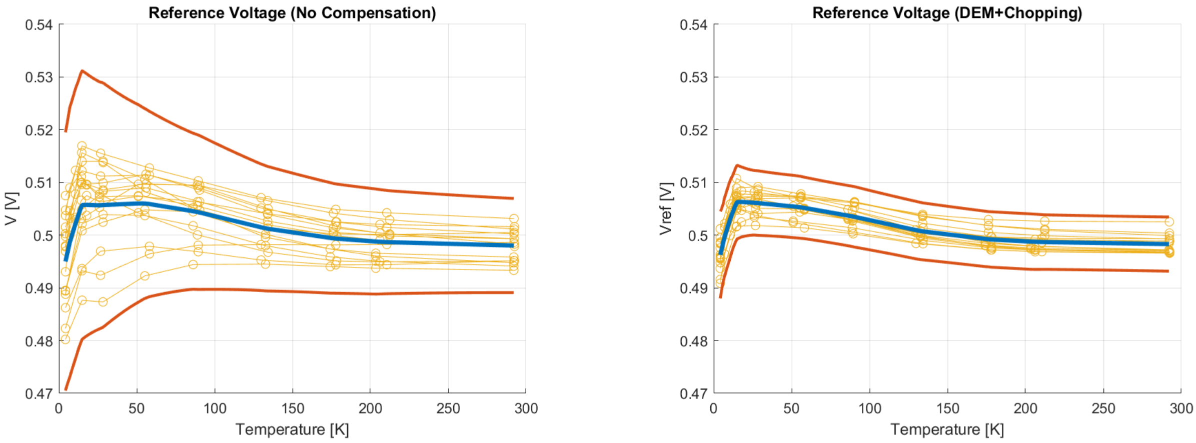
which can be seen as the ideal PTAT voltage with an additional offset term. At chopping phase 2, M_1 and M_2 are interchanged, resulting in a PTAT voltage equal to

$$V_{ptat,chop2} = -\delta V_{th} + n\frac{kT}{q} ln(p). \tag{3.17}$$

Mathematically speaking, by interchanging the core transistors, threshold voltage mismatch can be averaged to zero after averaging the PTAT voltage in both phases.

TC and variation after DEM and chopping

As mentioned in section 3.2, DEM and chopping were applied to the voltage reference in figure 3.1, although the switches are not shown in the figure. There are in total 16 samples of this architecture. The measurement results are shown in figure 3.8. At one temperature point, 32 arrangements were measured, based on 16 DEM phases and 2 chopping phases. An ideal averaging was performed in Matlab afterward. Figure 3.12 shows the measured reference voltage in the case of no compensation and the averaged reference voltage after performing chopping and DEM. Without compensation, mismatch at cryogenic temperatures is quite pronounced. DEM and chopping help to reduce the mismatch effectively, in which the 3σ reduces from 5.12 % to 3.37 % and 3.96 %, respectively. By doing DEM and chopping at the same time, 3σ of 1.65 % can be achieved. Meanwhile, TC also shows a significant improvement, reducing from 248.8 ppm/K to 134.2 ppm/K.



(a) Reference voltage without any compensation.

(b) Reference voltage after applying DEM and chopping.

Figure 3.12: Effect of DEM and chopping on the reference voltage.

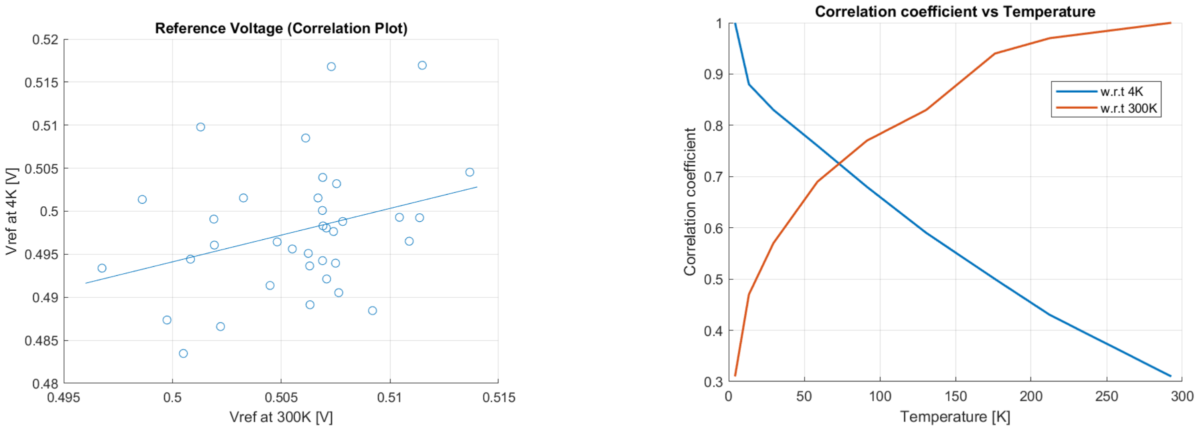
3.4.3. Trimming

Trimming involves measuring the output of the system, which contains the offset or gain errors, and then adjusting the value of certain components to bring the output back to the desired value [16]. Although trimming is costly in terms of the measuring effort, it is of rather low complexity compared with the other offset cancellation techniques. Furthermore, some errors such as the threshold voltage spread of the output transistor (M_6) can only be corrected for by trimming. Two types of trimming can be applied to this design to remove the PTAT- and offset errors introduced in section 3.4.1.

- **PTAT trim:** A PTAT trim aims to trim out the PTAT errors in the system. Since the first-order temperature coefficients in the PTAT- and CTAT voltage are set by the scaling factor (γ), it is possible to implement the PTAT trim by making γ (3.10) tunable.
- **Scaling trim:** A scaling trim can be used to scale up or scale down the reference voltage independent of temperature. It can be seen as providing another degree of freedom to adjust the overall gain setting of the system. By doing the scaling trim, the offset error in the reference can be removed.

Correlation between room temperature and cryogenic temperature

Trimming as mentioned above aims to correct the common static errors in the system. However, for a wide-temperature range voltage reference, performing trimming at a single point can only be very effective if the reference voltage at the edges of the temperature range has sufficient correlation. Figure 3.13a shows the measured reference voltage (untrimmed) at both 4 K and 300 K for each sample. The scattered points indicate that there is no strong correlation between the reference voltage at 4 K and 300 K. Figure 3.13b shows the correlation coefficient between the reference voltage at 4 K (and 300 K) and the reference voltage at other temperatures, calculated by the least square method. The correlation coefficient between 4 K and 300 K is only 0.31 while, it is 0.97 between 233 K and 300 K (−40 °C and 27 °C). This points out that it is likely that there is an effect appearing at cryogenic temperatures that reduces the correlation of the reference voltage with respect to the reference voltage at room temperature. In addition, 70 K is the temperature at which the reference voltage has an equal correlation with respect to 4 K and 300 K. This crossing point indicates that 70 K is the most effective temperature to apply a single-point trimming, in order to have maximum effect on both edges of the temperature range. Since 70 K is close to the liquid nitrogen temperature (77 K), performing trimming at this temperature is cheaper in terms of measuring cost.



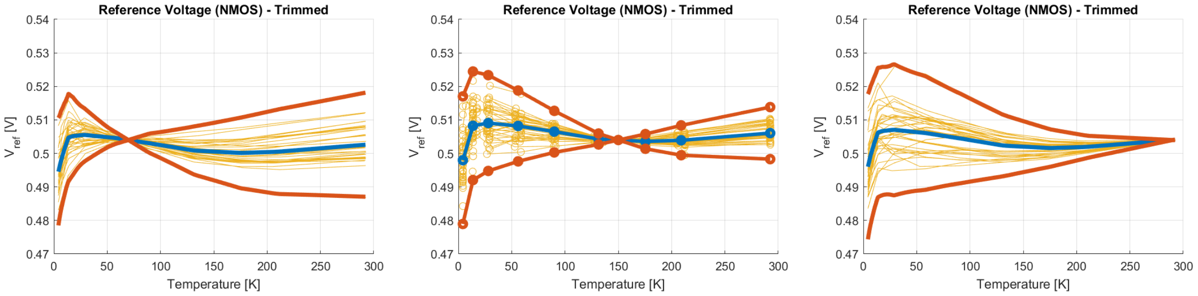
(a) Correlation between 300 K and 4 K.

(b) Correlation between each temperature.

Figure 3.13: Correlation between different temperatures.

TC and variation after trimming

As mentioned in 3.4.1, for MOS-based voltage references, threshold voltage spread is the major error contributor and it presents itself as a (first-order) offset error on V_{ref}. As a result, in the analysis in this section, the PTAT trim is used as a batch trim to correct the overall PTAT errors while the scaling trim is performed on each sample to remove the offset errors (the comparison of the effectiveness of the PTAT trim and scaling trim is shown in table 3.1 and appendix B). Since the architecture in figure 3.1 does not include additional elements to allow doing the scaling trim, the ideal scaling trim was done in Matlab to examine its effectiveness. Figure 3.14 shows the results after performing a single-point scaling trim for one batch of the references at three different temperatures. The most intuitive way for trimming is to perform it at the mid-temperature point, which is 150 K in this design (3.14b). Scaling trim effectively removed the offset errors, and therefore the TC and inaccuracy (3σ) both show 1.3 times improvement. Trimming at room temperature gives enormous benefits in terms of cost during measurements, while figure 3.13b suggests the optimal trimming point is at 70 K. Trimming at 300 K yields a TC of 292.8 ppm/K and trimming at 70 K a TC of 225 ppm/K, demonstrating that trimming at 70 K is indeed the most beneficial.



(a) Trimmed at 70 K.

(b) Trimmed at 150 K.

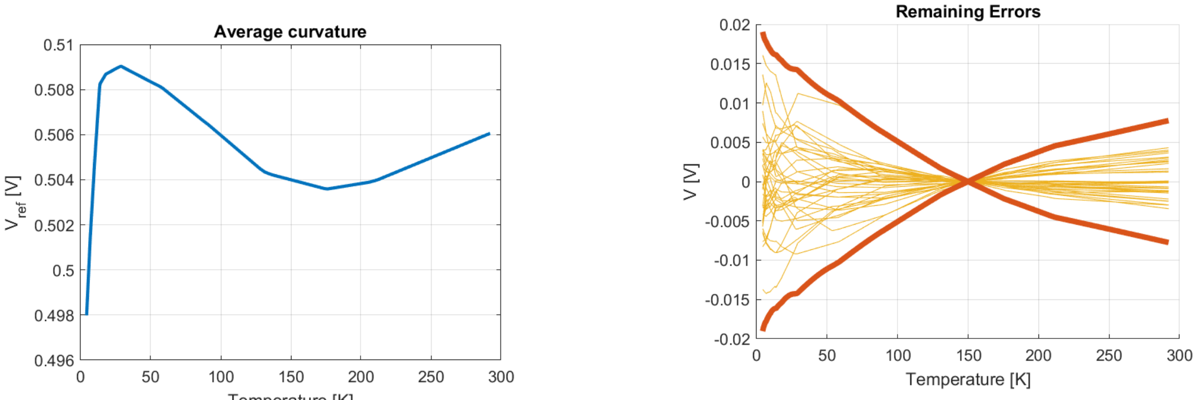
(c) Trimmed at 300 K.

Figure 3.14: Reference voltage after performing an ideal single-point scaling trim at 70 K, 150 K, and 300 K in Matlab.

3.4.4. Curvature correction

Besides mismatch and spread, there will be higher-order errors present in both the PTAT- and CTAT voltage as mentioned in section 3.3.3. To achieve a low temperature drift, curvature correction can be used to correct these higher-order errors. However, implementing error compensation requires an accurate device model valid over the full temperature range, which is currently not yet available. In order to determine the curvature that should be compensated for and what improvement in terms of TC can be achieved, an ideal curvature correction is performed in Matlab.

The approach for doing the ideal curvature correction is as follows: firstly, the average curvature is obtained from the trimmed reference voltage in figure 3.14b. The average curvature is shown in figure 3.15a. This curvature is then subtracted from all the reference curves. For this analysis, it is assumed that it is possible to design the compensation circuit which provides the exact trend of the average curvature.



(a) Average curvature.

(b) Remaining errors after curvature correction and a single point scaling trim.

Figure 3.15: Ideal curvature correction.

TC and variation after curvature compensation

The results after performing the curvature correction are shown in figure 3.15b. The TC before and after the curvature correction are 257 ppm/K and 210 ppm/K, respectively, calculated by using the box method 2.11. The main improvement comes from the fact that the kink at cryogenic temperatures is corrected now. However, the performance does not improve significantly by doing the curvature correction. TC=210 ppm/K is still far from the target specification, which is 10 ppm/K. The remaining errors at 4 K are 3.75 times larger than the remaining errors at 300 K. The unbalanced curves in figure 3.15b indicate that the mismatch below 100 K is limiting the performance.

3.4.5. Summary

This section aims to assess the effects of process variations on the reference voltage and quantify the performance improvement by using compensation techniques. The comparison table of the effects of different compensation techniques on TC and inaccuracy (3σ) is given in table 3.1. The analysis above points out the directions for designing a wide-temperature range voltage reference, which can be summarized as follows:

- There are two major obstacles when implementing high accuracy, and low TC voltage references, namely mismatch, and nonlinearity. Assuming the large curvature at cryogenic temperature can be corrected by using NMOS with a shorter channel length as core devices, the mismatch is the main bottleneck when implementing a reference that can work down to cryogenic temperatures instead of higher-order linearities.
- The effect of error sources on TC can be categorized into offset errors, PTAT errors, and nonlinear errors. Offset- and PTAT errors can be seen as gain errors in the system, and can therefore be removed by using either a PTAT- or scaling trimming.
- Neither a scaling trim nor a PTAT-trim can correct for random nonlinearities. To effectively remove the (statistical) nonlinearity introduced by the current sources and core transistors, DEM and chopping can be applied to the reference. Note that systematic nonlinearity can only be removed using a dedicated curvature correction scheme.

Given the above guidelines, the proposed design will implement several compensation techniques. The errors are expected to reduce to below 10 mV at 4 K after applying DEM, chopping, trimming, and assuming that the nonlinearities in both the PTAT- and CTAT voltage can be minimized (section 3.3.3). Figure 3.16 shows the remaining errors after performing the compensations techniques. As a result, a TC below 60 ppm/K can be expected. Compared with the work presented in [37], which focused on characterizing the sources of the error, and where the averaging was implemented in Matlab, the goal of this design is to have a full system where the averaging is done on-chip.

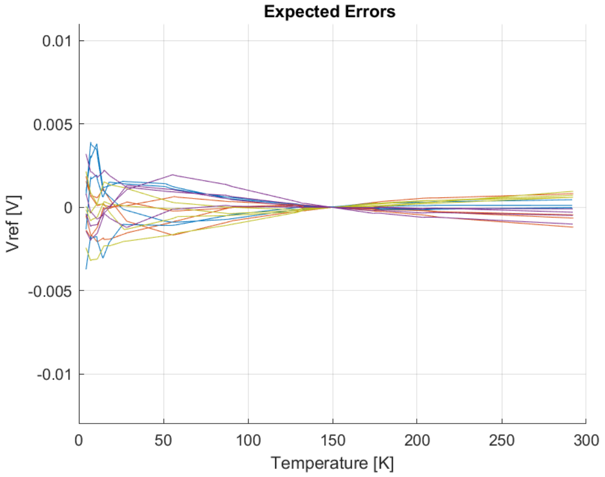


Figure 3.16: Residual errors after applying DEM, chopping, scaling trim, and curvature correction.

Table 3.1: Effects of different compensation techniques on TC and 3σ. Compensation techniques were applied to the NMOS-based reference, with the architecture features DEM. Note that the TC and 3σ are better than the results mentioned above. This is because there are only 16 samples for the DEM architecture, which is less than the basic architecture presented above (basic architecture refers to the architecture without switches for DEM. There are 36 samples for the basic architecture). Related figures can be found in appendix B.

	TC [ppm/K]	Inaccuracy (3σ) [%]	Figure
No compensation	254.7	5.12	B.1a
Single point scaling trim at 150K	217.46	4.02	B.2a
Single point PTAT trim at 150K	232.96	4.65	B.3a
DEM	196.89	3.37	B.1c
Chopping	216.9	3.96	B.1b
DEM + chopping	137.89	1.69	B.1d
DEM + chopping + scaling trim	110.89	1.19	B.2d
DEM + chopping + PTAT trim	123.36	1.39	B.3d
Curvature correction	214.47	5.12	B.4a
Curvature correction + scaling trim	156.58	4.02	B.4b
DEM + chopping + scaling trim + curvature correction	53.07	1.19	3.16

4

Circuit Design

This chapter presents the design of the proposed voltage reference at both the system level and the transistor level. Firstly, the system overview will be presented. Secondly, each sub-block in the system will be introduced, including the architecture choice, circuit implementation, and simulation results. Finally, the layout of the full chip will be presented.

4.1. System Overview

From the analysis in chapter 3, it is clear that mismatch is the main bottleneck for designing high-accuracy cryogenic voltage references. To achieve the target specifications, several offset-cancellation techniques, such as dynamic element matching (DEM), chopping, and trimming are implemented in the design. However, this will introduce ripples in the reference voltage. In order to obtain a clean reference voltage, these ripples must be removed.

The overall system is shown in figure 4.1, which consists of a reference generator, averaging circuit, and clock generator. The reference generator first generates the temperature-independent voltage V_{ref}. The variation in V_{ref} is up-converted to higher frequencies by applying DEM and chopping. After this, a switched-capacitor integrator will be used to remove these up-converted mismatch errors. In terms of functionality, the integrator can be effectively seen as a low-pass filter. In principle, it samples each DEM phase, integrating them to generate a constant output V_{out}. Based on a single clock input, all the required dynamic control signals for the system are generated on-chip by the clock generator. In addition, there is an auxiliary circuit block for static configuration and multiplexing for characterization purposes.

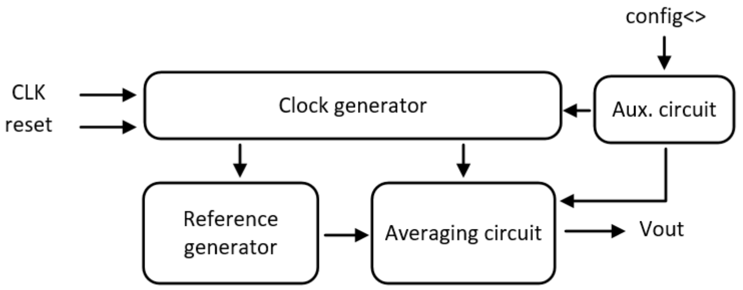


Figure 4.1: System overview.

4.2. Reference Generator

This section presents the architecture of the reference generator, which contains the reference core and the compensation circuit. The architecture is the same as presented in [6, 37] but with slightly different sizing. As mentioned in chapter 3, in order to investigate whether channel length is indeed the cause of the kink in PTAT voltage at cryogenic temperatures, two versions of reference generator with different core device sizing are implemented. To make a fair comparison, only the sizing of the core devices has been changed. More details of the error budgeting and the sizing considerations of the reference generator can be found in [37]. The following sub-section presents the architecture, implementation of compensation techniques, and simulation results.

4.2.1. Reference core

Architecture

The architecture of the reference generator is shown in figure 4.2. As introduced in section 3.2, by biasing M_1 and M_2 in weak inversion with different current densities, the difference in their gate-source voltage (ΔV_{gs}) shows a PTAT behaviour. This voltage is then converted into current by R_{ptat} and is mirrored to the output branch by M_3 and M_5. The PTAT voltage generated across R_{out} can be expressed as

$$V_{ptat} = \frac{\Delta V_{gs}}{R_{ptat}} \cdot m \cdot R_{out} = n\frac{KT}{q} ln(p) \cdot \frac{mR_{out}}{R_{ptat}}, \tag{4.1}$$

where m is the current ratio between M_5 and M_3, and p is the current ratio between M_4 and M_3. V_{gs6} has a CTAT behaviour as mentioned in the previous chapters. By choosing the proper scaling factor mR_{out}/R_{ptat}, a first-order temperature-independent reference voltage can be obtained. It can be expressed as

$$V_{ref} = n\frac{KT}{q} ln(p) \cdot \frac{mR_{out}}{R_{ptat}} + V_{gs6}. \tag{4.2}$$

The current ratios p and m depend on the current carried by the current sources M_{3-5}. Due to the finite output impedance, the accuracy of p and m will be affected. To alleviate this error, a feedback loop formed by M_7 and M_8 is added. Given that V_{gs2}=V_{gs7}, I_{D2} is equal to I_{D7}. Consequently, since $(W/L)_{M8}$=$(W/L)_{M4}$, V_{gs4} and V_{gs8} are equal, resulting in V_{ds3}=V_{ds4}. With the drain voltage of M_3 and M_4 being equal, the circuit is more robust against supply changes, hence improving the line regulation and accuracy. Note that this feedback loop does not ensure the drain voltage of M_5 is the same as M_3 and M_4. Therefore, M_5 is the limiting factor for the line regulation. Note that using cascodes to improve the supple rejection is non-trivial in this design due to the increased V_{gs} at cryogenic temperatures and the large variation in the biasing current as mentioned in section 2.4. The circuit is stable as the negative feedback loop is dominant. This can be seen from follows: assuming there is an increased signal at the drain of M4, this increased signal will be translated to a decreased signal at the gate of M8 due to the gate-drain inversion of M3. This signal will be inversed two more through of M8 and M2. Ending up as a decreased signal at the drain of M4, which indicates that the negative loop is presented.

Sizing considerations

As the biasing current in this architecture has a PTAT behaviour, the current reduces as the temperature reduces. At cryogenic temperatures, the current sources might enter weak inversion due to the increased threshold voltage and reduced current. The W/L ratio should be chosen to make sure the current sources stay in deep strong inversion at room temperatures. Accordingly, M_{3-5} have a W/L ratio of 1 µm/7.5 µm. Drain current mismatch can be modeled as

$$\sigma^2_{\Delta I_D/\overline{I_D}} = \sigma^2_{\Delta\beta/\overline{\beta}} + \left(\frac{g_m}{\overline{I_D}}\right)^2 \sigma^2_{\Delta V_{th}}, \tag{4.3}$$

where $\sigma_{\Delta V_{th}}$ is the variability of threshold voltage = $A_{V_{th}}/\sqrt{WL}$ and $\sigma_{\Delta\beta}$ is the variability of the current factor = $A_\beta/\sqrt{WL}$ [27]. Thus, it is beneficial to use a large transistor with a small transconductance to reduce the drain current mismatch. L=7.5 µm is used for the current sources. Long transistors not only reduce mismatch because of larger geometry but also provide a larger V_{gs} to improve the drain current matching [35]. Further

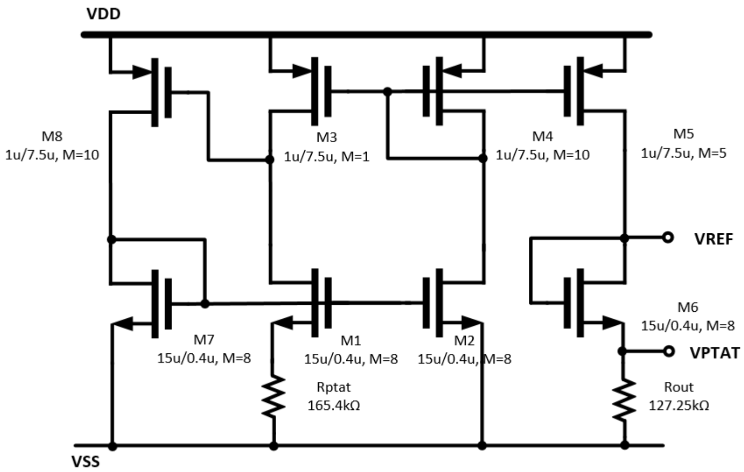


Figure 4.2: Architecture and sizing of the voltage generator.

increasing the width might not help, as a larger width increases the g_m as well.

To generate the PTAT voltage, the core devices M_1 and M_2 must be in weak inversion. Working in weak inversion usually required devices to have a low-current level. Since the biasing current has a PTAT behaviour, the core devices have a high W/L ratio to ensure they operate in weak inversion at the highest operating temperature. L=0.4 µm is chosen for the experiment to investigate if kink depends on channel length as mentioned in chapter 3.

The total current budget of the reference generator is decided to be 2 µA at 4 K, so that more budget can be left for other blocks. Since the bias current shows a PTAT behaviour and is expected to reduce 5x at 4 K, the total current budget at 300 K is 10 µA. For the current ratio in the current sources being $M_8 : M_3 : M_4 : M_5$=10 : 1 : 10 : 5, the current level at the unit size transistor M_3 can be calculated as

$$I_{D,M3} = \frac{10\mu}{10+1+10+5} = 384nA. \tag{4.4}$$

The required R_{ptat} can be obtained based on Ohm's law,

$$R_{ptat} = \frac{\Delta V_{gs}}{I_{D,M3}} = 169.2k\Omega, \tag{4.5}$$

where ΔV_{gs} =65 mV based on the simulation. The actual value of the R_{ptat}=165.4 kΩ is decided together with the layout considerations. R_{out}=127.2 kΩ is chosen such that the first order temperature dependence in V_{ref} (equation 4.2) can be cancelled.

4.2.2. Offset cancellation techniques

The offset cancellation techniques implemented in this design include dynamic element matching, chopping, and trimming. The compensated architecture is shown in figure 4.3. This is the extended architecture of figure 4.2.

Dynamic element matching (DEM)

The basic idea of DEM is to swap the position of identical elements in a circuit. By swapping the instances at a given frequency, the mismatch error is up-converted to the higher frequencies and can be filtered out by averaging. As the PTAT current in the output branch can be approximated by $m \cdot nKT/qln(p)$, errors in ratio p and m will therefore be directly translated into an error in the reference voltage. As a result, DEM is applied to the current sources M_{3-5} in this design. With the ratio of $M_3 : M_4 : M_5 = 1 : 10 : 5$, there are in total 16 unit current sources. Ideally, random sampling from all the possible arrangements for the current sources gives the best result in terms of mismatch reduction. However, this increases the complexity of implementing the

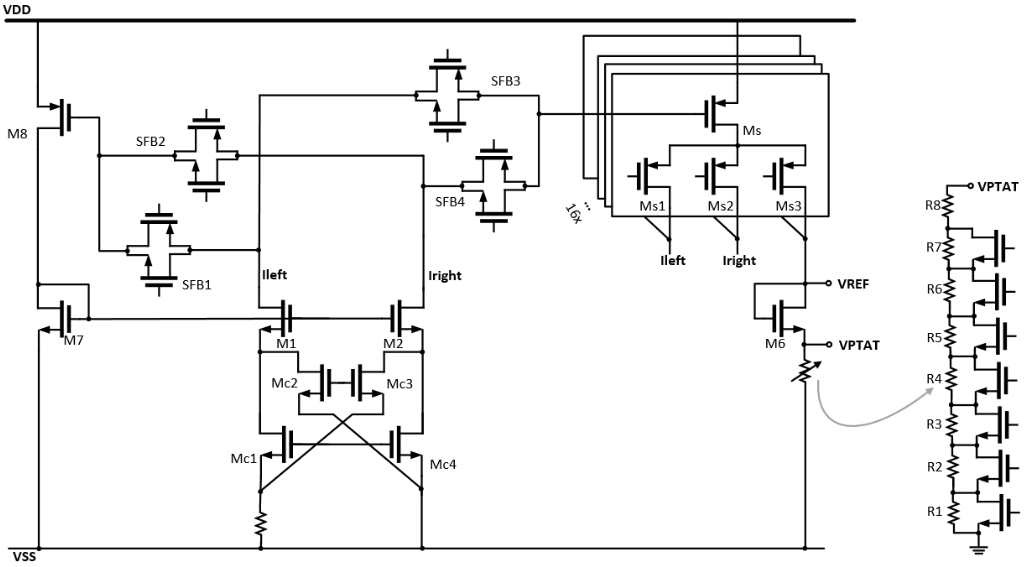


Figure 4.3: Reference generator with the switches for offset compensation.

digital control logic. Since the unit current source (in the branch of M_3) is the main error contributor, 3σ inaccuracy on V_{ref} can already be reduced by six times by letting all of the 16 current sources be in this position at least once based on the simulation (table 5.1). For the implementation, the current sources are shifted by one position every phase. By using 16 different arrangements, the errors caused by the current ratio mismatch will be alleviated. Note that the idea of DEM is to average the errors instead of fully canceling them, implying that there will still be some higher order errors present [16].

The switches that are used to implement the DEM (M_{s1-s3}) are implemented with PMOS devices. Due to subthreshold leakage current, both the PTAT- and CTAT voltage at room temperatures are slightly lower in this architecture (figure 4.3) compared with the basic architecture (figure 4.2). The ideal current ratio p is 10, which means $p = I_{right}/I_{left} = 10$. However, the actual current ratio is approximated as

$$p' = \frac{I_{right} + 6I_{leakage}}{I_{left} + 15I_{leakage}} < 10, \tag{4.6}$$

where $I_{leakage}$ represents the leakage current for the switches that are opened but still connect to the branch. Since the subthreshold leakage is expected to be less pronounced at lower temperatures, no further leakage-reduce technique is adopted in this design.

Chopping

Compared with DEM, chopping first up-modulates the input signal and then de-modulates it back to DC. While the desired signal is de-modulated, the offset is up-modulated to the chopping frequency (f_{chop}) [16]. Chopping is applied to the core transistors M_1 and M_2 in this design. It is fundamentally the same as doing DEM on core devices M_1 and M_2 as only the mismatch is up-converted but not the signal. By doing so, the threshold voltage- and beta mismatch between the two core devices can be averaged out.

Transistors M_{c1-c4} are used to interchange the position of M_1 and M_2. Between the two chopping phases, the current ratio needs to be interchanged to maintain stability. Switches that are used to interchange the current ratio are implemented with passgates. The on-resistance of the passgates strongly depends on the threshold voltage of the devices. Having a closer look at the gate of M_8 shows that the node voltage can be approximated by $V_{DD} - V_{gs8}$, which is expected to vary between 500 mV to 410 mV based on available measurement data. Further reducing the on-resistance of the passgates would require to use a larger width. However, this introduces a gate leakage current flowing to the left branch (I_{left}) at the same time, thereby reducing the accuracy. Consequently, the on-resistance of these switches at 4 K is one of the limiting factors that prevent using higher chopping frequencies. The chopping frequency that will be used in this design is in the range of 300 Hz to 2.4 kHz

16 arrangements of the current sources are implemented during one chopping phase. Given the 2 possible chopping states, 32 possible arrangements of the core transistors and current sources can be used.

Trimming

This design uses trimming to correct the static error. As mentioned in section 3.4.1, certain errors show a PTAT behaviour, which allows to remove them by using a PTAT trim. The scaling ratio mR_{out}/R_{ptat} controls the slope of the PTAT voltage. R_{out} is made tuneable to allow such a PTAT trim. Since trimming is expensive in terms of time and cost during measurements, it will only be used as batch trim in this design. That is, all the references from the same batch will have the same configuration for the PTAT trim.

The resolution of the trimming network is decided to be 1 mV. The fine resolution allows for minimizing the PTAT errors further. However, the linearity in the PTAT voltage is limited by the kink as presented in figure 3.3. Other nonlinearities are negligible compared with the kink, therefore the trimming network does not need to have a fine resolution. The expected PTAT error based on the measurement data is 23 mV, for the resolution of 1 mV, 5-bit is required. In order to also cover unexpected changes at cryogenic temperatures, the range of the actual trimming network has been extended. The implementation of the trimming network is shown in the right-most part of figure 4.3. The resistors used in this trimming network R_{1-7} are binary weighted while R_8 is the static resistor. With R_{out} in figure 4.2 equals 127.2 kΩ and PTAT voltage equals 250.8 mV at 27 °C, the R_{LSB} (R_1) can be calculated as follows,

$$R_{LSB} = \frac{V_{LSB}}{V_{PTAT} \times R_{out}} = 507.23\Omega. \tag{4.7}$$

The switches used in the trimming network is implemented with NMOS devices as they only need to switch low voltage ($<V_{ref}$). Sizing considerations of the switches have been made such that the error introduced by on-resistance and finite-off resistance of the switches is negligible [37].

4.2.3. Start-up transistor

This architecture requires a start-up circuit to push the circuit out of the zero-current state. At room temperature, starting up the circuit can be done even without a startup circuit due to the subthreshold leakage currents. However, when the temperature decreases, the subthreshold leakage also decreases, making it less straightforward to start up the circuit. Therefore, a startup transistor is required to let the current flow. In the design presented in [6] the startup circuit consists of a comparator, which determines whether the circuit is in the on- or off-state, and generates a start-up pulse to trigger the start-up transistor if needed. However, due to the architecture and operation of the averaging circuit (which will be covered in the following section), the output of the reference generator will change from 0 to around 500 mV from phase to phase. Therefore, this approach is not suitable for this design.

The start-up transistor is shown in figure 4.4, marked in grey. This start-up transistor is controlled by the reset pulse for the clock generator. When the system is powered on, a pulse (1.1 V) will be applied to the gate of this start-up transistor. Consequently, the gate of the M_8 will be pulled down to the ground, the drain of M_{7},8 and the gate of $M_{1,2}$ will be pulled up due to the gate-drain inversion of M_8. Current is therefore allowed to flow through all the branches. After the circuit starts up and reaches its desired state, the start-up transistor will be connected to a logic 0, hence being invisible to the rest of the circuit.

4.2.4. Simulation results

This section presents the simulation results of the reference generator. The results presented here are based on the architecture shown in figure 4.3. Figure 4.5 shows the reference voltage (V_{ref}) and PTAT voltage (V_{ptat}) from −40 °C to 27 °C from both schematic and parasitic extracted (PEX) simulations. With a nominal V_{ref}=486 mV, the temperature coefficient in the schematic level is 23 ppm/K. After parasitic extraction, a voltage shift of 5 mV is observed in V_{ref}, while V_{ptat} is almost unchanged. Since the error in V_{ptat} after

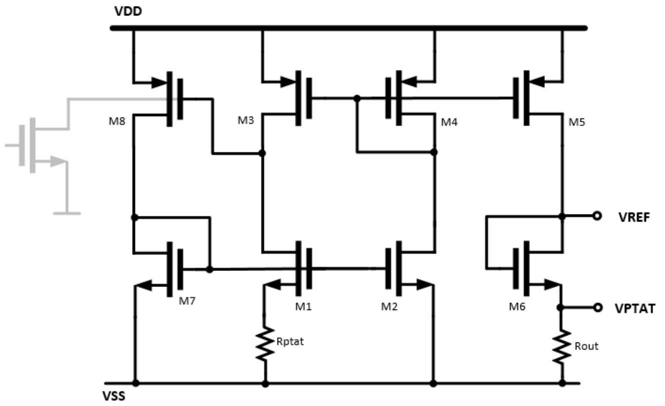
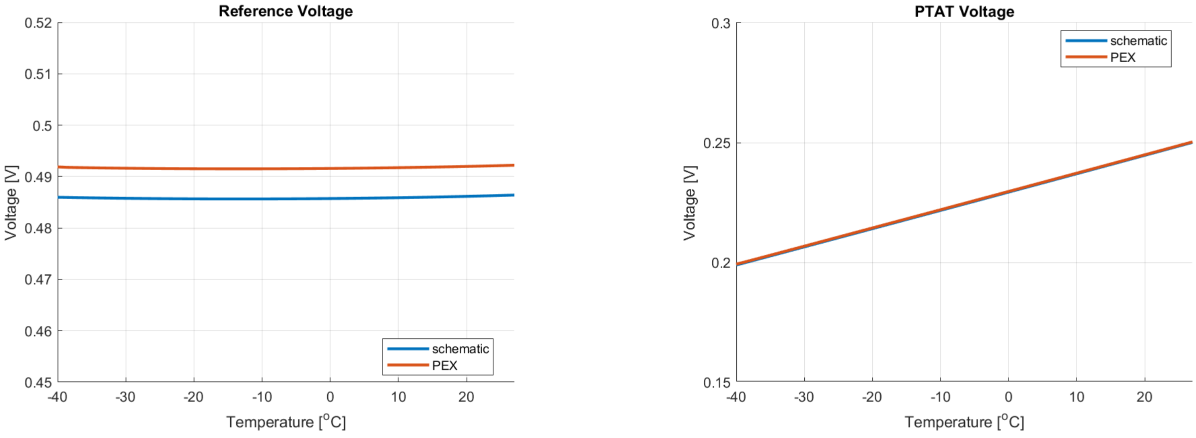


Figure 4.4: Architecture of the reference generator with start-up transistor. The start-up transistor is marked in grey.

parasitic extraction is below 200 µV, the change in V_{ref} is mainly from the CTAT voltage as can be seen from equation 4.2. This threshold shift can be attributed to the second-order effect caused by layout. Figure 4.6 shows the range and resolution of the PTAT trimming network. The blue line in the figure 4.6 is the simulation result with the mid-code setting while the other lines represent various codes.



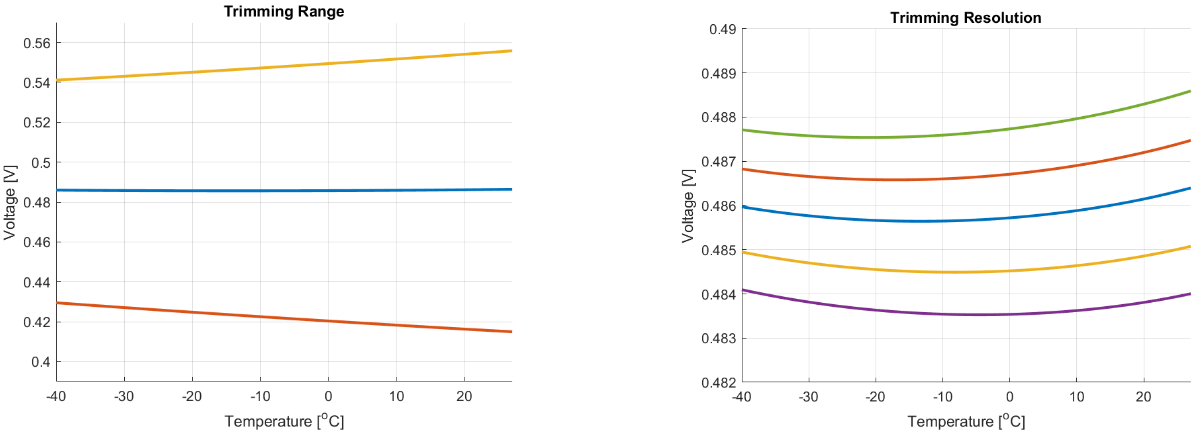
(a) Reference voltage vs. temperature.

(b) PTAT voltage vs. temperature.

Figure 4.5: Reference and PTAT voltage vs. temperature.

Figure 4.7 shows the effects of process spread and mismatch on the reference voltage. Figure 4.7a shows the results of 200 runs of a Monte Carlo simulation at the nominal corner (TT). The temperature coefficient and the accuracy are 916.3 ppm/K and 2.96 %, respectively, without performing trimming. Based on the simulations, the threshold voltage spread within a corner at room temperature is expected to be smaller than 3 mV. The main source of variation in V_{ref} is caused by the current source mismatch. Figure 4.7b shows the reference voltage at different corners. Since MOS-based references are fundamentally extracting the threshold voltage of the transistor (M_6 in this design), the simulation results are consistent with the expectation that V_{ref} is lower when NMOS is in the fast corner (FS/FF) and higher when NMOS is in the slow corner (SF/SS). The threshold voltage spread is around 50 mV across corners, which is the main limiting factor for implementing batch-independent voltage references by using MOS transistors.

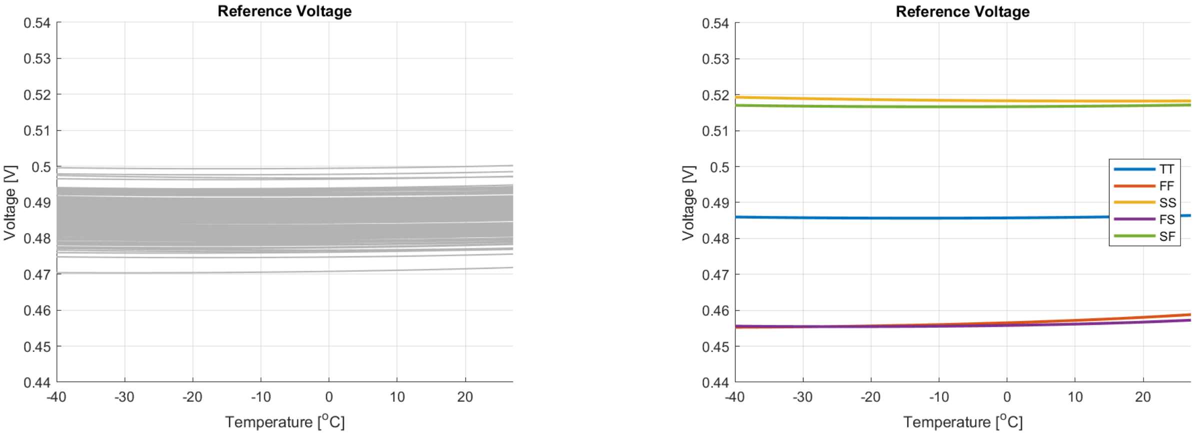
Figure 4.8 shows the reference voltage in time domain at 27 °C. The simulation result is obtained by running a Monte Carlo simulation for 1 sample, with the DEM frequency set at 1 kHz. The moments when swapping the devices are indicated by arrows in the figure. Different arrangements of the current sources produce different output voltages, which explains why the square-wave-like ripples are observed. Each little square wave represents a different current source mismatch. The threshold voltage mismatch between M_1 and M_2 can be



(a) Trimming range.

(b) Trimming resolution.

Figure 4.6: Range and resolution of the PTAT trimming network.



(a) The effects of process variations on the reference voltage.

(b) The effects of process spread on the reference voltage.

Figure 4.7: Reference voltage versus temperature.

observed from the overall voltage shift between the two chopping phases. It is added on top of the current source mismatch. The ripple is approximately 16 mV, which is significantly limiting the accuracy and should therefore be removed.

The simulation results of line regulation and noise simulation at 27 °C are shown in figure 4.9. The achieved line regulation is 15 mV/V with a nominal supply of 1.1 V. The integrated noise from 1 Hz to 10 Hz is 45 µV$_{\text{rms}}$. The majority of the noise is from the flicker noise of the unit size current source M_3, which contributes 88.3 % of the total noise. Since the reference generator is biased with a PTAT current, the power consumption is expected to reduce at cryogenic temperatures. The current consumption reduced from 9.58 µA at 27 °C to 7.59 µA at −40 °C.

4.3. Offset-Compensated SC-Integrator

One of the drawbacks of DEM and chopping is that it generates ripples, as mismatch is up-modulated to the harmonics of f_{dem}. This ripple is expected to be 18 mV at 4 K based on the measurement data, which is larger than what can be tolerated based on the specifications. Therefore, a circuit that performs the averaging is required. This section presents the system-level design of the averaging circuit, including the motivation, architecture choice, transistor-level implementation, and simulation results.

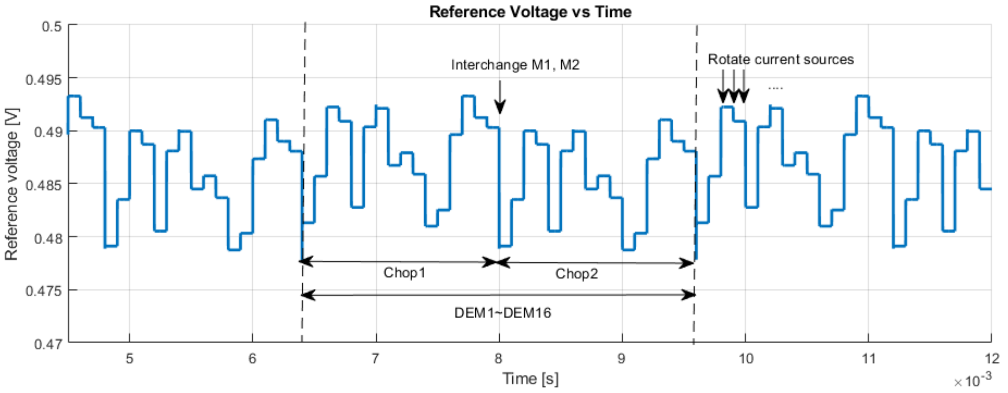
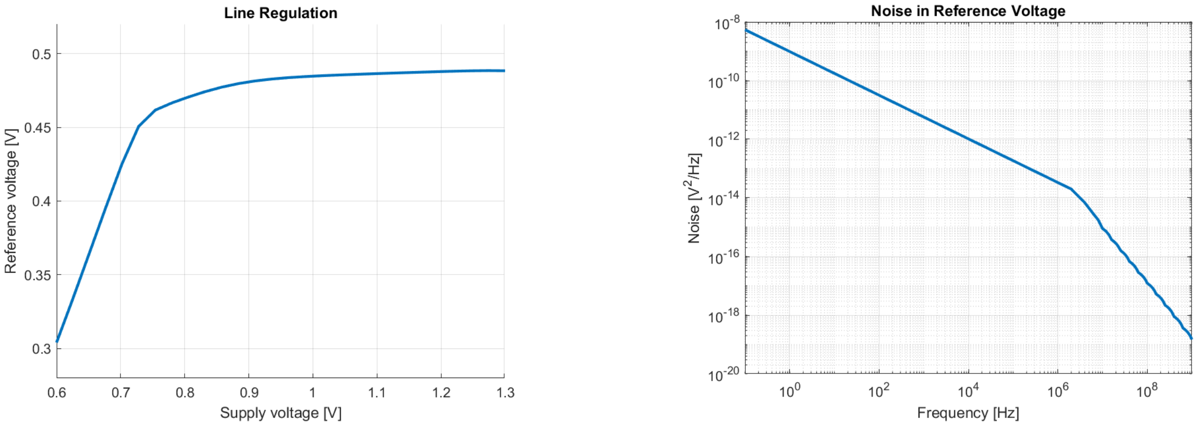


Figure 4.8: Reference voltage in the time domain when applying DEM and chopping at 27 °C.



(a) Line regulation.

(b) Noise.

Figure 4.9: Line regulation and noise simulations at 27 °C.

4.3.1. Motivation for using SC-Integrator to do the averaging

The most intuitive way to suppress the ripple is to add a low-pass filter (LPF) at the output of the system to filter out all the high-frequency noise caused by DEM and chopping [38]. However, the reference generator proposed by [37] is designed for characterization purposes rather than for implementing dynamic offset cancellation techniques. In order to investigate if the kink is related to the channel length of the core devices, most of the sizing in the reference generator remains unchanged with respect to the design presented in [37]. Large transistors and low current levels of the reference generator limit the operation speed. Low chopping frequencies make using an RC-LPF impractical as it requires a large silicon area. Although a switch-capacitor filter (SC-filter) can be used to replace the bulky resistor, it usually requires a buffer at the output to generate a buffered reference voltage. This is not desired as the accuracy of the voltage reference will then depend on the performance of the buffer. More compensation techniques have to be applied to alleviate the errors introduced by the buffer, which further increase the complexity. Many start-of-the-art implementing SC-notch filter for ripple reduction [19]. Through this approach, mismatch will first be integrated into triangular-like ripples, and then be sampled at the zero crossing point. However, due to the architecture of the reference core itself, this is not feasible in the design.

Switched-capacitor circuits are another possibility for implementing the averaging. As introduced in chapter 2, some of the SC-based voltage references store the PTAT- and CTAT voltage on a capacitor during the sampling phase, and then sums them together during the holding phase [39–41]. However, accurate coefficients for summing operations are not easily implemented by using passive SC networks, especially if there are many DEM phases. The same idea can be extended to use a SC-integrator for performing the summing operation. In this approach, each DEM phase will be sampled and then integrated. The active integration provides a possibility to simplify the SC network as only one summing coefficient is required.

Among these options, the SC-integrator is chosen in this design for performing the averaging. Compared to a buffer, offset and gain compensation are easier to implement in the integrator. Furthermore, the capacitors that are used in the integrator feedback network can be used as another degree of freedom for trimming. It provides a possibility to trim out the offset between samples caused by the threshold voltage spread.

4.3.2. Working principle

As mentioned in the introduction above, the integrator is chosen to do the averaging. A correlated double-sampling (CDS) integrator has been chosen for performing the offset and gain compensation. The architecture is shown in figure 4.10 [42]. As illustrated in figure 4.10a, during the sampling phase, the sampling capacitor (C_s) samples the input voltage V_{ref} and at the same time, C_h samples the offset voltage of the opamp. Assuming the opamp has infinite DC gain, the voltage on these two capacitors can be expressed as

$$V_{cs} = V_{ref} - V_{cm}, \tag{4.8}$$

$$V_{ch} = V_{-} - V_{cm} = V_{cm} + V_{os} - V_{cm} = V_{os}, \tag{4.9}$$

where V_{cm} is the input common-mode voltage of the opamp and V_{os} is the offset voltage of the opamp.

During the integration phase (figure 4.10b), S_1, S_3, S_5 are opened and S_2, S_4 are closed. C_s will be discharged to

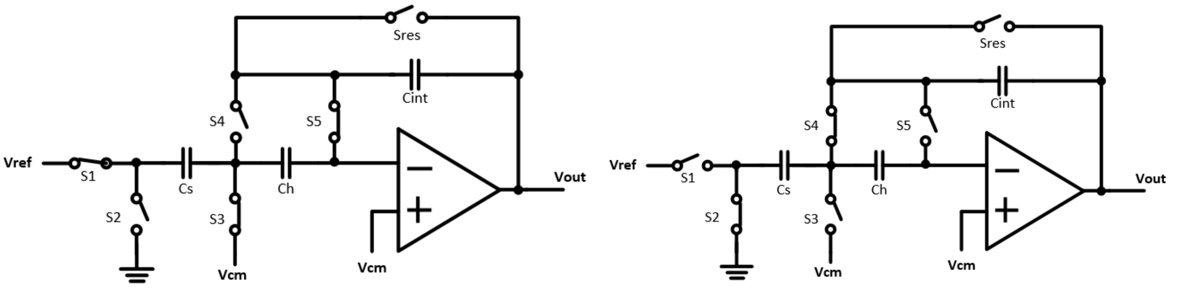
$$V_{cs} = 0 - V_{cm}. \tag{4.10}$$

The charge sampled on the C_s has therefore been transferred to the integration capacitor (C_{int}), causing a voltage change ΔV_{cint} that can be described as

$$\Delta Q = \Delta V \times C_s = [V_{ref} - V_{cm} - (0 - V_{cm})] \times C_s = V_{ref} \times C_s, \tag{4.11}$$

$$\Delta V_{cint} = \frac{\Delta Q}{C_{int}} = V_{ref} \times \frac{C_s}{C_{int}}. \tag{4.12}$$

Without the presence of C_h, the left plate of C_{int} is approximate V_{os} during the integration phase, resulting in incomplete integration. Through the help of C_h, the offset voltage can be subtracted. Since the integrator cannot distinguish the difference between offset and low-frequency noise below the sampling frequency, the CDS technique also helps to reduce 1/f noise.



(a) Sampling phase (phase1).

(b) Integration phase (phase2).

Figure 4.10: Architecture of the CDS integrator.

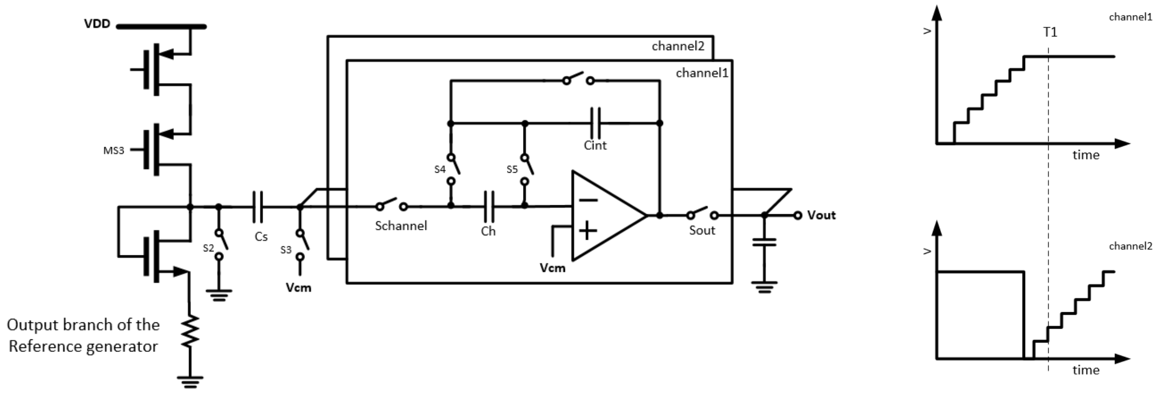
The final output voltage after the whole integration period can be approximated as

$$V_{out} = V_{cm} + n \times \Delta V_{cint} = V_{cm} + V_{ref} \times \frac{C_s}{C_{int}}, \tag{4.13}$$

where n is the number of DEM and chopping phases (integration cycles). The output voltage depends on V_{cm} and the ratio between C_s and C_{int}, and it is an increasing ramp during the integration. This comes with two advantages: firstly, it provides a non-inverting output. Secondly, the output voltage does not change dramatically from phase to phase, which makes the settling requirements of the opamp more relaxed. Note that the equations above neglect the finite gain error introduced by the opamp. In reality, the finite gain of the opamp degrades the integration accuracy. Extended derivations regarding finite gain can be found in [43].

4.3.3. Modified architecture

During the integrated period, the output is an increasing ramp and it is not ready for being used as a reference voltage. In order to make the output voltage continuously available, two channels are connected together to work in a ping-pong mode. This operation is illustrated in figure 4.11b. The first integrator starts integrating from phase 1 to phase 32. After full integration, it enters the hold mode, after which it is connected to the output node (at time T_1). At the same time, the other integrator is integrating. The continuous output voltage is obtained by interchanging these two integrators.



(a) Modified architecture.

(b) Output waveform.

Figure 4.11: Modified architecture and time domain waveform illustration.

Having one integrator in the hold mode implies that there is always one integrator that does not require to perform sampling and integrating actions. As a result, two integrators can share the same sampling capacitor (C_s) as shown in figure 4.11a. Sharing the same C_s not only saves the area but also improves the matching between two channels at the same time. $S_{channel}$ and S_{out} are clocked with anti-phase. They are used to select which integrator is connected to the output (hold mode) and which integrator is connected to the sampling capacitor (integration mode). One of the drawbacks of two-channel operation comes from the switching spikes when switching between the output of the two integrators. To reduce this error, a capacitor is added at the output to filter out these spikes.

Note that V_{ref} is generated by the reference generator, and it is expected to be around 470 mV to 520 mV depending on the threshold voltage. For the switch S_1 in figure 4.10, that means the V_{gs} of this switch might be as low as 0.5 V if using a 1.1 V supply. Figure 4.12 shows the measured on-resistance of the minimum size NMOS and PMOS device at 4 K. The data is characterized under V_{ds} =10 mV, by applying a fixed voltage (1.1 V) at the gate and sweeping the common-mode voltage (V_{cm}) for the switches. When the input voltage is at mid-rail (=0.55 V), R_{on} might be so large that it affects the settling. Note that figure 4.12 only shows the measurement result of one sample. Due to threshold voltage spread and mismatch, the actual on-resistance may vary significantly over samples, implying that this is not a robust solution for cryogenic temperatures. It is reported that the threshold voltage of the minimum size NMOS device might be larger than 0.6 V at 4 K in this technology [24], which could potentially cause settling issues if used for this purpose.

Using a larger device helps to reduce the on-resistance of the switch. However, it comes at the cost of leakage and charges injection which is not desired. To avoid this issue, it is possible to remove S_1 by adjusting the timing. The modified architecture is shown in figure 4.11a, where M_{s3} is the cascode switch that is used for DEM as shown in figure 4.3. The timing diagram is shown in figure 4.13. During the sampling phase, S_3 and S_5 are closed. Current flows through the output branch of the reference generator and charges C_s with respect to V_{cm}. After this period, there is a short non-overlapping time to ensure S_2 and S_3 will not be closed at the same time, avoiding losing the charge stored on C_s. As for the integration phase, S_2 is grounded, and S_4 is closed, causing a voltage change on C_{int}. Meanwhile, M_{s3} is opened to cut off the current at the output branch of the reference generator. Otherwise, the non-zero on-resistance of S_2 will cause a voltage drop, which is lifting the ground of the left plate of C_s when there is too much current flowing through.

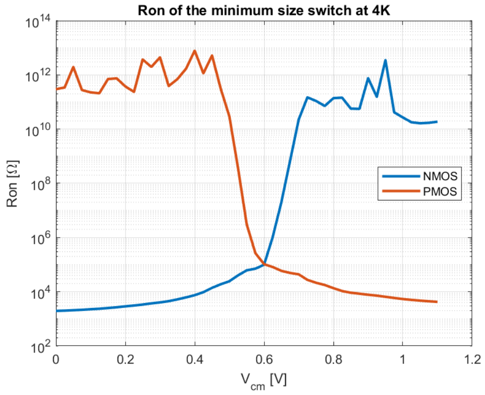


Figure 4.12: Measured on-resistance of the minimum size (W/L=120 nm/40 nm) NMOS/PMOS devices at 4 K.

In the conventional SC-integrator operation, there is also a non-overlapping time after the sampling phase to avoid that S_1 and S_4 in figure 4.11a are closed at the same time to let V_{ref} charge C_{int}. This should be taken care of in this design as well. However, in the modified architecture, where S_1 is removed, having this non-overlapping time might potentially cause another problem. At the moment when S_2 and S_4 are opened, the left plate of C_s observes a voltage jump from 0V to around 500 mV. Assuming that all the switches are open (in the conventional non-overlapping operation), this sudden jump might affect the output voltage through the input of the opamp. To avoid this, the timing for the SC-integrator is modified from the dashed line in grey to the black solid line in figure 4.13. After the charge is transferred to C_{int}, S_3 closes first before S_2 opens to avoid the output being affected by this sudden jump. Since the integrator is working at a low frequency in the range of 40 kHz to 80 kHz, a slight reduction in the sampling time will not introduce any significant errors.

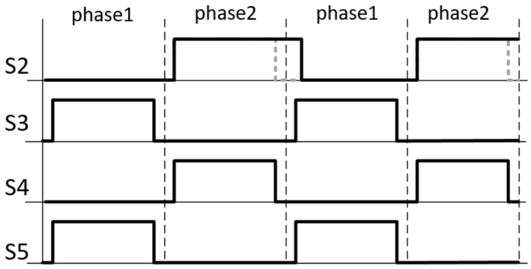


Figure 4.13: Timing for the operation of the SC-integrator. Phase1 is the sampling phase and phase2 is the integration phase.

4.3.4. Non-idealities in the SC-integrator

This section introduces the nonidealities in the SC-integrator, which is used as input for the real circuit implementation.

Charge injection and clock feedthrough

When switches turn off, the charge in the channel will be injected to the surrounding circuits through the source and drain terminals. The amount of this charge can be approximated as $Q_{ch} = WLC_{ox}(V_{gs} - V_{th})$. This charge might result in an error if it deposits on the capacitor. Charge injection is proportional to the device size, V_{gs}, and V_{th}. Since most of the switches in the design have an almost constant input and fixed gate voltage, the charge injection error would result in an offset error. However, there will be some residual higher-order errors present. The effect of charge injection is expected to be worse at room temperatures as the threshold voltage is lower when the temperature is higher.

Besides charge injection, the switching transitions in the clock might also introduce errors via the coupling from the gate-source or gate-drain capacitance, this is usually referred to as clock feedthrough. The errors caused by the clock feedthrough can be approximated as $\Delta V = V_{CLK} W C_{overlap}/(W C_{overlap} + C_H)$, where $C_{overlap}$ is the overlap capacitance per unit width and C_H is the hold capacitance [34].

Offset and mismatch

Since the integration step depends on the ratio of C_s and C_{int} as pointed out in equation 4.12, the mismatch in these two capacitors will affect the final output voltage directly in the form of a gain error. Besides capacitance mismatch, the mismatch between the input pair of the opamp will introduce offset and it will be integrated into the output voltage if there is no offset compensation being performed.

Leakage

There are several effects that cause leakage currents, such as subthreshold leakage, substrate leakage, and gate leakage. Although they are from different effects, the influence that they cause on this design can be summarized in the following two ways,

- **Reduced accuracy:** If the charge is leaking away from the capacitor during each sampling phase, then the final output voltage will be slightly different from the ideal value. Since subthreshold leakage is more problematic at high temperatures and is expected to disappear at cryogenic temperatures, such a temperature-dependent error will translate to curvature in the output voltage.

- **Degraded noise:** During the hold mode, if the charge keeps leaking away, then a voltage drop will be observed. This voltage drop over time will translate to ripples present in the output voltage, degrading the noise. Due to the long holding time, this might be problematic when the integrator is in hold mode. As long as the voltage drop caused by gate leakage is below the noise level, it will not be a concern.

The errors caused by leakage are proportional to the hold time and inversely proportional to the hold capacitor size. Since this design will be working at a low frequency, extra care should therefore be taken to avoid leakage problems.

4.3.5. Switches

The implementation of switches is shown in figure 4.14. All the switches in the SC-integrator, S_{2-5}, are implemented with NMOS devices, as they only need to switch low voltage V_{cm} (this will be covered in section 4.3.7). As mentioned in section 4.3.4, there will be nonidealities introduced by the switches, such as charge injection and clock feed-through. Pass-gates are often used to alleviate the effect of the charge injection [34]. However, they come at the cost of slightly increased complexity. Given that the pass voltage of the switch is far away from the mid-V_{DD}, the additional complexity does not outweigh the lower resistance. Dummy switches can also help to reduce the charge injection and clock feed-through [34]. Nevertheless, adding more devices might potentially cause more gate leakage currents. It is therefore decided to use only a single NMOS device.

Since the chopping frequency that will be used is in the range of 312.5 Hz to 2.5 kHz, C_{int} has to hold the output for a long time. To reduce gate leakage, S_4 is driven by a lower gate voltage of 0.9 V instead of 1.1 V. To provide more freedom at the measurement phase, this is controlled by a tuneable supply. Consequently, S_4 is implemented as a low threshold voltage (LVT) transistor. $S_{2,3,5}$ are implemented with standard threshold voltage (SVT) transistors to reduce the subthreshold leakage.

As mentioned in section 4.3.3, S_{select} and S_{out} are added for two-channel operation. Since the switching transients in the integrating channel might couple to the holding channel, these two switches are implemented as a T-switch with its middle node of the switch connected to ground [9]. Two switches in series are clocked with the same phase while the third switch is clocked in the anti-phase. However, there is still a residual error from the parasitic capacitance of the T-switch. To reduce this residual error, S_{select} and S_{out} are all minimum-size devices. The reset switch is also implemented as a T-switch, to avoid subthreshold leakage adding nonlinear errors to the output voltage. The third switch connects the middle node of the series switches to V_{cm}. Since the left plate of C_{int} is approximate V_{cm}, the off-resistance of the switches can be increased in this way by making the potential of drain and source equal [41].

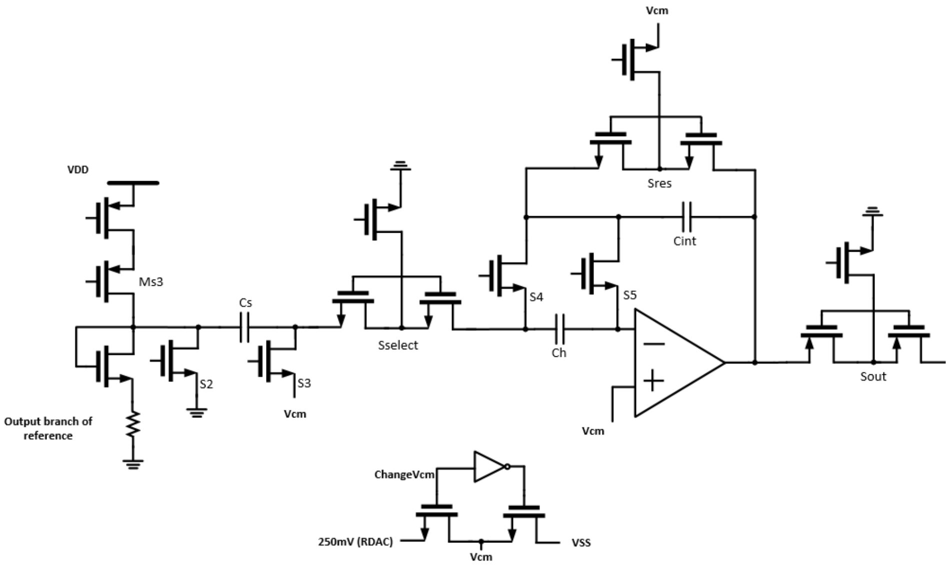


Figure 4.14: Switches implementation.

4.3.6. CDAC

The ratio between C_s and C_{int} determines the integration step as shown in equation 4.12. For the target output voltage of 900 mV, the input voltage given by the reference generator ≈ 486 mV. With a total of 32 phases, the capacitance ratio is decided to be 0.054 with C_s=1.08 pF and C_{int}=19.8 pF. The actual sizing of the capacitance is usually determined by the noise specification. However, in this design, the value is determined by the allowable leakage, as leakage is the dominant source of error rather than noise. By using such a large capacitor, the noise and charge injection can also be suppressed.

Due to process variations, the output voltage of the two integrators will be slightly different from each other. This difference will translate to the final output voltage as low-frequency noise (at the frequency when switching between two channels). Therefore, C_{int} is implemented as a tunable capacitor as shown in figure 4.15, so that the mismatch between the two channels can be trimmed out. The other functionality of C_{int} is to provide an additional degree of freedom required for the scaling trim. As mentioned in chapter 2, threshold voltage spread is the main bottleneck for a CMOS-based reference to achieve performance comparable to a BJT-based reference. Whereas the reference generator does not easily allow for a scaling trim, the integrators in the averaging circuits can be conveniently implemented with a tunable gain by making the integration capacitor tunable. Tuning this gain can therefore serve as the required scaling trim.

Since there are two types of reference generators on-chip (which will be covered later), the trimming range is designed to cover 150 mV, which is larger than required for the reference generator introduced in section 4.2.1. The CDAC is designed to be able to trim out all the threshold voltage spread for both versions of the reference generator. The resolution is determined by the target accuracy, which is 0.5 mV. In order to cope with the uncertainty in the output range of the opamp at cryogenic temperatures (as it will be explained in section 4.4.3), C_9=2.9 pF in figure 4.15 has been added to allow the reference to be shifted down by an additional 100 mV. Having this option ensures that the opamp can have sufficient gain, even when the output range at cryogenic temperatures is lower than expected.

C_{int} consists of two parts, a static capacitor (C_x) and the CDAC. The capacitors in the CDAC are binary weighted with C_{LSB}=10 fF. They are implemented using unit-size capacitors to optimize matching. The switches used in the CDAC are thick-oxide NMOS devices. In case of thin-oxide switches, gate leakage from the nine switches introduces a significant error. Based on simulations, the threshold voltage of the thick-oxide devices at room temperature is comparable to the threshold voltage of the thin-oxide standard-Vth devices at room temperature. However, it is not clear if the threshold voltage of the thick-oxide devices increases in a similar way as thin-oxide devices. To avoid the on-resistance being too large at 4 K, the gate of these switches is driven by a level shifter, providing a possibility to increase the gate voltage in case the 1.1 V supply is not sufficient. In addition, the middle node of these switches is connected to the same source of V_{cm} to alleviate the subthreshold leakage [41].

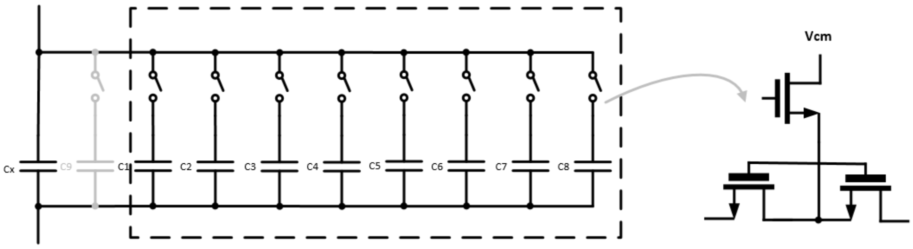


Figure 4.15: CDAC and switches implementation.

4.3.7. Input common-mode voltage of the opamp

Due to the single-ended structure of the integrator, the output voltage depends on V_{cm} as pointed out in equation 4.13. Ideally, using V_{cm}=0 can address this issue. However, designing an opamp that can handle the exact rail-to-rail output voltage from 0 V to 1.1 V while maintaining sufficient DC gain is non-trivial. $V_{cm} = 250mV$ is therefore chosen for ease of integration based on the output range that opamp can support. To get rid of the dependence of V_{cm}, V_{cm} will be changed from 250 mV to ground during the half of the integration period. The time domain waveform illustration is shown in figure 4.16. By doing so, the output voltage does not depend on the quality of V_{cm} anymore.

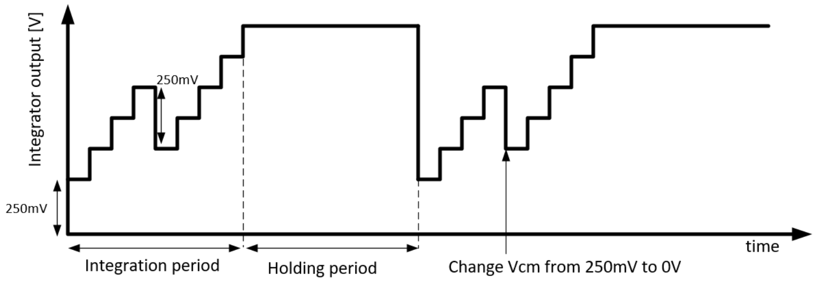
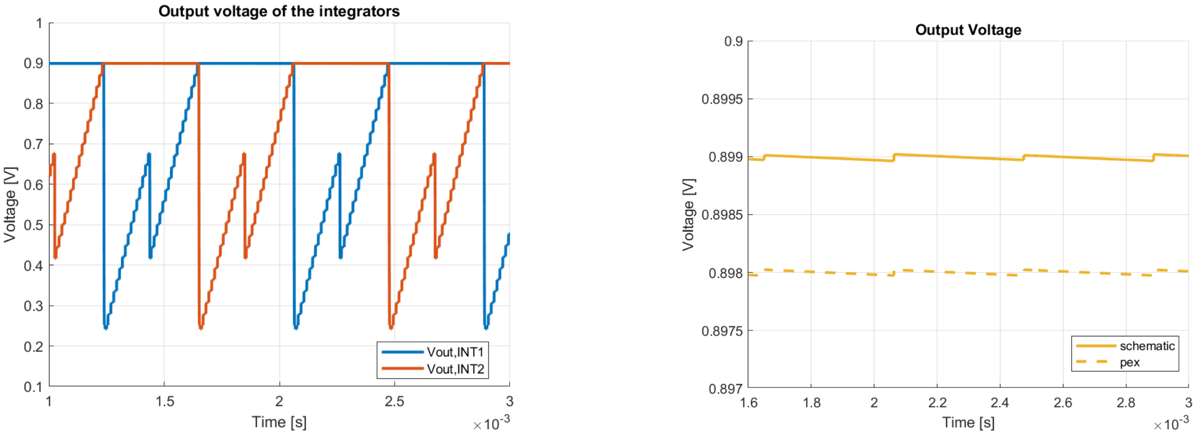


Figure 4.16: Illustration of changing V_{cm}.

4.3.8. Simulation results

Figure 4.17 shows the simulation results of the averaging circuit. Figure 4.17a shows the output waveform of the two integrators. The jump after half of the integration period is due to the change V_{cm}. Having two integrators working in a ping-pong mode, the continuous output can be obtained. The voltage droop (47 µV) observed in V_{out} is caused by the gate leakage of the switches. The extracted simulation result is shown in the dashed line in 4.17b. Due to the parasitic capacitance, the actual C_{int} is larger than in the schematic. The error caused by the parasitic capacitance can be seen as a constant gain error and can be trimmed out by applying a scaling trim.

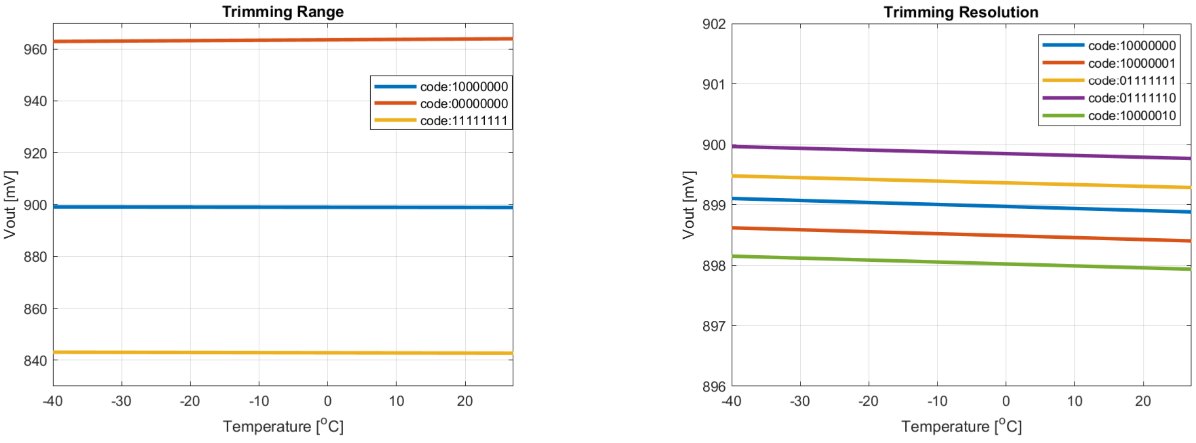
Figure 4.18 shows the trimming range and the trimming resolution of the CDAC. The blue line in the figure is the simulation result with the mid-code setting. Note that there is a slight increase in voltage from 27 °C down to −40 °C. The difference is due to the subthreshold leakage at high temperatures, coming from S_2 and S_5.



(a) Output waveform of the integrators.

(b) Output waveform of the averaging circuit.

Figure 4.17: Output waveform of the integrators and the averaging circuit at −40 °C.



(a) Trimming range.

(b) Trimming resolution.

Figure 4.18: Trimming range and resolution of the CDAC.

4.4. OpAmp

The opamp is the core building block for the integrator. This section first covers the specifications of the opamp, and then presents the architecture, sizing considerations and simulation results.

4.4.1. Specifications

Nonidealities from the opamp, such as finite gain, bandwidth, offset and noise affect the integrator and hence degrade the overall performance of the voltage reference. For implementing an accurate voltage reference, there are mainly two concerns.

- **Performance should be stable over temperature:** Achieving performance that is stable over temperature is taken as the first priority. Otherwise, the temperature-dependent errors introduced by the opamp will degrade the TC of the reference voltage.
- **Accurate summing procedure:** the main task for the integrator is to sample all DEM phases and then sum them up together. Since the integration is an increasing ramp from 250 mV to 900 mV, DC gain of the opamp should be as constant as possible for the whole output range to allow for an accurate summing procedure. The output swing is therefore considered to be a highly important specification.

These bring the requirements of the opamp, listed in table 4.1. To satisfy the above-mentioned requirements, the opamp must have sufficient DC gain over the full output range from 250 mV to 900 mV. These two specifications are therefore considered the first priority. Noise degrades the resolution of the voltage reference, however, the drift over time is expected to be limited by the gate leakage of switches as mentioned in section

4.3.5 and the mismatch between two channels. As a result, no hard requirement on the noise is listed.

Since the residual errors (mismatch and temperature drift) from the reference generator are expected to cause 60 ppm/K temperature drift as shown in figure 3.16, the error allowed for the opamp is decided to be 2 ppm/K (0.5 mV).

Table 4.1: Opapm specifications.

Spec	Value
DC gain	49.5 dB
BW	1.8 MHz (for f_s =80 kHz)
Input common-mode	0 and 250 mV
Output range	250 mV-900 mV

4.4.2. Architecture

Gain-enhanced current mirror OTA

A current mirror OTA is chosen for the opamp architecture, as it can provide a large output swing while also being power efficient. The architecture is shown in figure 4.19a, which consists of a differential pair and three current mirrors. Assuming M_6 and M_4 have a W/L ratio of B, then the current generated by the input pair is mirrored to the output branch by a factor of B. As a result, the DC gain of this architecture can be calculated as

$$A = g_{m1} \times B \times r_{o6}. \tag{4.14}$$

Assuming the transistors are in strong inversion, the transconductance of M_1 (g_{m1}) and the output impedance of M_6 (r_{o6}) can be expressed as [44]

$$g_{m1} = \frac{2I_{bias}}{V_{gs1} - V_{th1}}, \tag{4.15}$$

and

$$r_{o6} = \frac{1}{\lambda B I_{bias}}, \tag{4.16}$$

where I_{bias} is the bias current and λ is the channel-length modulation factor. Therefore, the initial gain A that the simple current mirror OTA can provide can be approximated as

$$A = \frac{2}{(V_{gs1} - V_{th1})\lambda}. \tag{4.17}$$

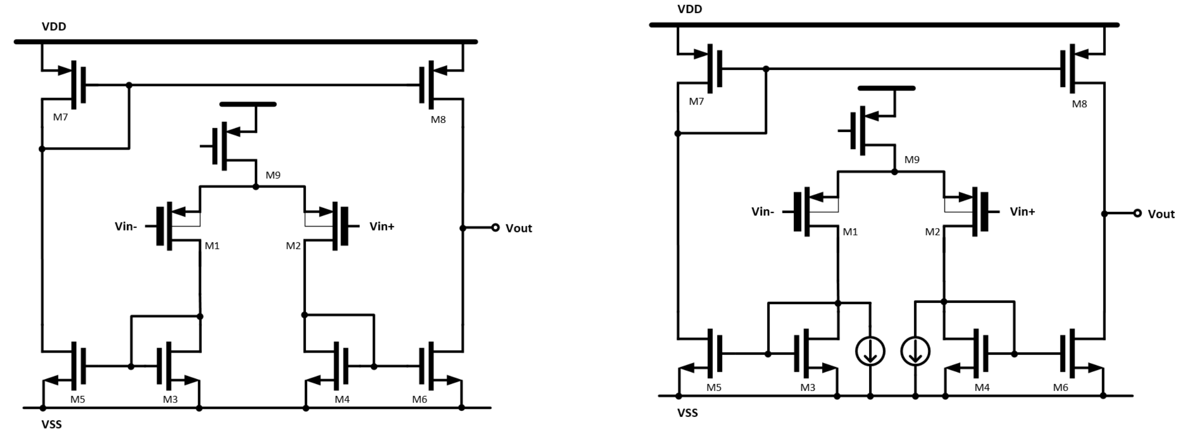
In advanced technologies, a single transistor usually cannot provide sufficient DC gain. Note that increasing B does not help to increase the DC gain (as shown in equation 4.17) because it reduces the output impedance by the same factor. Some topologies that have been commonly used for gain boosting are cascode- and multi-stage amplifiers. However, this comes at the cost of reduced output swing and higher power consumption and is therefore not desired. Another way to increase the DC gain is to use the gain enhancement technique as proposed in [45, 46]. This is done by placing an additional two current sources beside M_3 and M_4 as shown in figure 4.19b. These two current sources are intended to draw some current away from M_3 and M_4, making less current being mirrored to the output branch. Effectively, it can be seen as putting more current to the input pair to increase g_{m1} while maintaining the same output impedance. Assuming that these two current sources can draw k times the bias current (kI_{bias}) away, then the current flowing through the output branch becomes

$$I_{D6} = (I_{bias} - kI_{bias}) \times B = (1-k) \times BI_{bias}. \tag{4.18}$$

Consequently, the DC gain of the amplifier can now be expressed as

$$A_{GE} = \frac{2I_{bias}}{V_{gs1} - V_{th1}} \times B \times \frac{1}{\lambda(1-k) \times BI_{bias}} = \frac{1}{1-k} \frac{2}{(V_{gs1} - V_{th1})\lambda}, \tag{4.19}$$

which has now been increased by a factor of $1/1-k$. k can be well defined at the design stage by biasing the current sources properly. Note that increasing k comes at the cost of decreased bandwidth, therefore a trade-off has to be made here [45].



(a) Conventional current mirror OTA.

(b) Gain-enhanced current mirror OTA.

Figure 4.19: Architecture of the conventional and gain-enhanced current mirror OTA.

Allow using V_{cm}=0

As mentioned in section 4.3.7, the output voltage of the integrator is approximated as $V_{cm} + V_{cint}$. To make the output voltage independent of V_{cm}, V_{cm} has to be zero. However, V_{cm} has to be large enough to keep the input pair of the opamp in saturation. Given the condition for maintaining the transistor in saturation: $V_{ds} \geq V_{gs} - V_{th}$, this requires V_{cm}>V_{D1}-V_{th}, where V_{D1} is the drain voltage of M_1. To achieve this, either V_{th} has to be higher or V_{D1} has to be lower. Using higher V_{th} is not desired, as there might not be sufficient headroom at 4 K due to the low supply voltage (1.1 V) and increased threshold voltage. Therefore, the preferred way would be to lower V_{D1}, such that 0 can be used as the input common-mode voltage even for small V_{th}.

In order to allow for this, the architecture is modified from figure 4.19b to figure 4.20. The main difference is that the current mirrors have been changed into active current mirrors. By using the active current mirror, the drain voltage of M_1 and M_2 can be shifted down to the designed value [47]. The value of $V_{D1,2}$ (V_{ls} in the figure) is determined to be 250 mV. As 250 mV is large enough to keep $M_{3-4,10-11}$ in saturation and also low enough to keep the input pair in saturation. The auxiliary amplifiers are implemented with a simple 5OTA. The DC gain of the auxiliary amplifiers is not so critical as V_{ls} does not need to be very accurate. V_{ls} is generated on-chip by the resistive divider. Looking into this auxiliary amplifier and M_4, it is essentially a two-stage amplifier connected in a unity-gain buffer configuration. Compensation is therefore required to make this loop stable. In addition, an output capacitor (C_{Load}) is added to make the overall opamp stable.

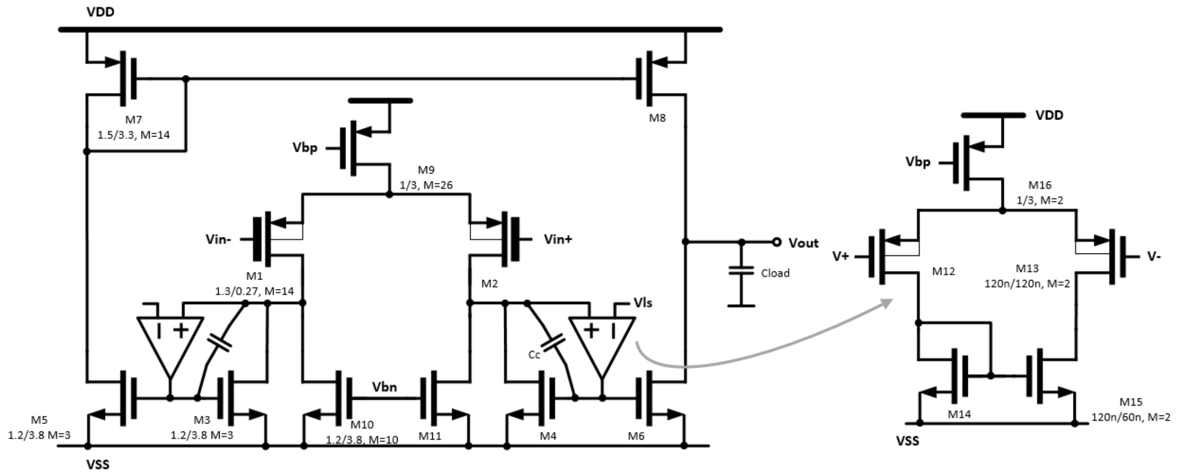


Figure 4.20: Opamp architecture and sizing.

4.4.3. Sizing considerations

The sizing of the devices is labeled in figure 4.20, unless otherwise mentioned, the unit is given in μm. Sizing is only labeled for half of the circuit as the designed opamp is symmetric. The reasoning for the sizing will be given in this section.

To handle the low input voltage, PMOS devices are used for the input pair. In addition, they are implemented with thick-oxide devices to reduce gate leakage. Input pair $M_{1,2}$ is biased in weak inversion with a PTAT current. The I_D-V_{gs} relation when the device is in weak inversion can be approximated as equation 2.6. Therefore, the transconductance can be derived as

$$g_m = \frac{\partial I_D}{\partial V_{gs}} = \frac{I_D}{nkT/q}. \tag{4.20}$$

By biasing with a PTAT current, which is proportional to nkT/q, the temperature dependence in g_m can be canceled. As a result, bandwidth can remain the same for the entire operating temperature range. To avoid the threshold voltage of the input pair increasing too much at 4 K, which would limit the available headroom for the current source M_9, the bulk terminals of the input pair are connected to the source terminal to lower the threshold voltage.

Given that V_{dsat} of M_9 is around 100 mV at −40 °C and V_{cm} is 250 mV for the first half of the integration period, the available V_{ds} for M_9 is limited by the threshold voltage of the input pair. When going to cryogenic temperatures, this threshold voltage will increase, and hence reduce the available headroom for M_9. Due to the lack of cryogenic device models (especially for thick-oxide devices), there is a potential risk of M_9 being pushed into the linear region when the threshold voltage increases more than expected. The opamp was designed to handle an expected threshold voltage increase of 250 mV. However, to eliminate the uncertainty in the estimation, M_9 has a supply different from the other parts of the opamp. In case the threshold voltage increase is larger than expected, it remains possible to tune the supply such that M_9 can be brought back into saturation.

Transistors that formed the current mirrors have a long length and are biased with a small current to increase the output impedance. Usually, current sources are sized to have large V_{gt} for better current matching and low-noise performance. However, this reduces the output swing at the same time, which is not desired in this design. Therefore, current sources are not in deep strong inversion.

The gain-bandwidth product of the opamp (figure 4.19a) can be approximated as

$$GBW = B \times \frac{g_{m1}}{2\pi C_L}, \tag{4.21}$$

where C_L is the load capacitance. Using larger B gives better speed, but since the error introduced by the incomplete settling is to first order temperature-independent (as bandwidth is fixed), it can be seen as an offset error. Consequently, bandwidth is not a major concern in this design. Having a larger B not only consumes more power but also introduces larger parasitic capacitance, reducing the phase margin. In this design, $B = 1$ is chosen. The current sources that are used for gain enhancement are formed by M_{10} and M_{11}. They are biased in strong inversion and have the same unit-size W/L with current sources M_{3-6}. The ratio between M_4 and M_{11} is 3:10. In principle, the sizing of M_{10} and M_{11} does not need to be the same as the other current sources. The same unit-size transistor is used for simplicity and to avoid higher-order errors.

As for the auxiliary amplifier, the bias current is kept small to make the difference of the g_m of the first and the second stage amplifier (auxiliary amplifier and M_4) larger, so that the stability of this loop can be compensated more easily. An additional compensation capacitor C_c=829 fF is added to make sure the loop has sufficient phase margin. However, this compensation capacitor together with the large M_{10}, M_{11} adding significant parasitic at the gate of M_4 and M_6. Hence introduce the non-dominate pole. A large C_L=5.8 pF is therefore needed to maintain the overall stability.

4.4.4. Biasing network

The required biasing current for the opamp is generated on-chip by the PTAT generator as shown in figure 4.21. The PTAT current is generated in the same way as illustrated in section 4.2.1. This current is mirrored out to the opamp core through M_{b3} and M_{b7}, where M_{b3} is used to bias the gain-enhanced current sources $M_{10,11}$ and M_{b7} is used to bias the current sources for the main opamp and auxiliary opamp. Since the opamp is biased with PTAT current, the power consumption is expected to reduce five times at cryogenic temperature.

To avoid the coupling between the two integrator channels, two biasing networks are used. In order to provide more freedom in the measurement phase, the biasing network is adapted to allow for external biasing as well. These modifications are drawn in grey in figure 4.21. The switches for enabling the external biasing have the size of 1 µm/ 0.4 µm to reduce the on-resistance and have the same multiplier factor as the current source (M_{b4-6}) to make the layout easier. V_{bext} is connected to the bondpad directly.

Similar to the reference generator, this biasing circuit also requires a start-up network. The startup transistor is implemented with an NMOS device, with its drain terminal connected to the gate of M_{b4}. The start-up pulse that triggered this start-up transistor is the same as for the reference generator.

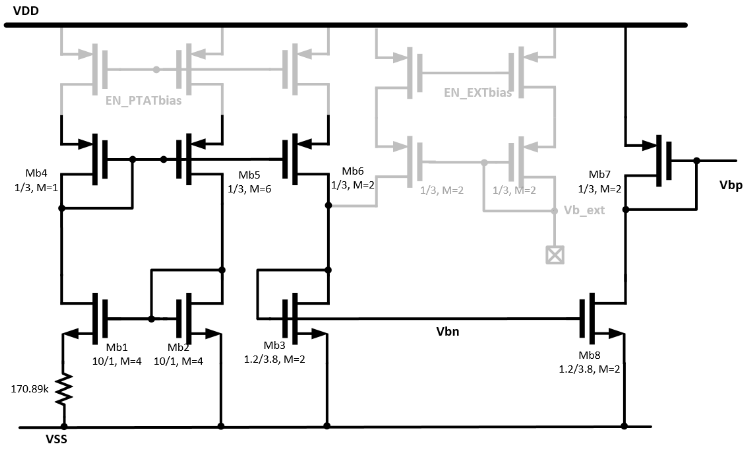


Figure 4.21: Biasing network.

4.4.5. Simulation results

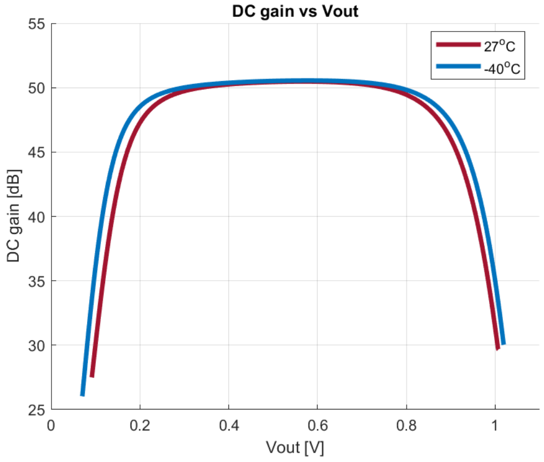


Figure 4.22: DC gain vs. output voltage

Figure 4.22 shows the DC gain versus output voltage at both 27 °C and −40 °C. The output swing is larger at −40 °C because the opamp is biased with the PTAT current. Given that the drain can be approximated as $I_D = \beta/2(V_{gs} - V_{th})^2$, to first order, a reduction in current reduces the required headroom of the transistors. The output voltage of the integrator is low at the first few integration steps, in case the opamp cannot provide sufficient output swing, V_{cm} is generated by 3-bit RDAC to give more freedom during measurement.

Figure 4.23 shows the opamp performance across different corners. The opamp is simulated in a closed-loop configuration with the output voltage of approximately 250 mV, which is the worst-case scenario in terms of DC gain. The discrepancies between corners are mainly due to the resistance spread in the biasing network. At the FF corner, the resistor has the smallest value, which results in a larger biasing current. Consequently, in the FF corner, the bandwidth is larger due to the increased g_m and the DC gain is lower due to the decreased output impedance. In the SS corner, this will be the exact opposite case.

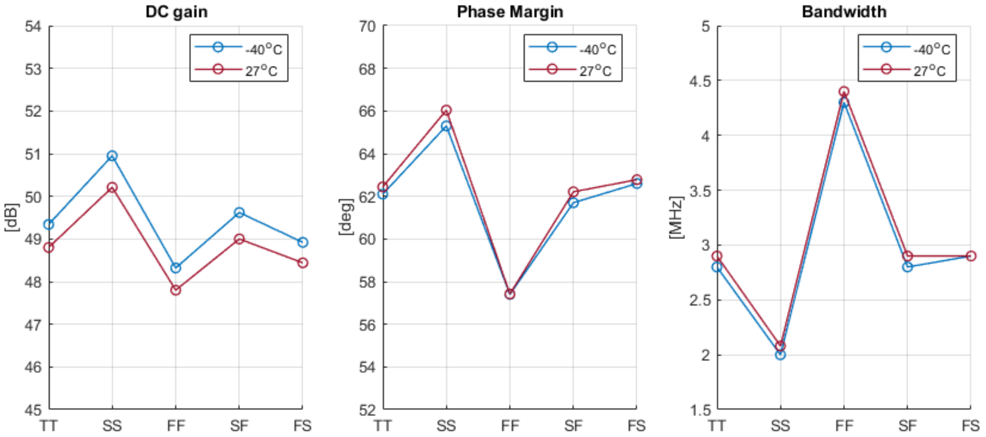
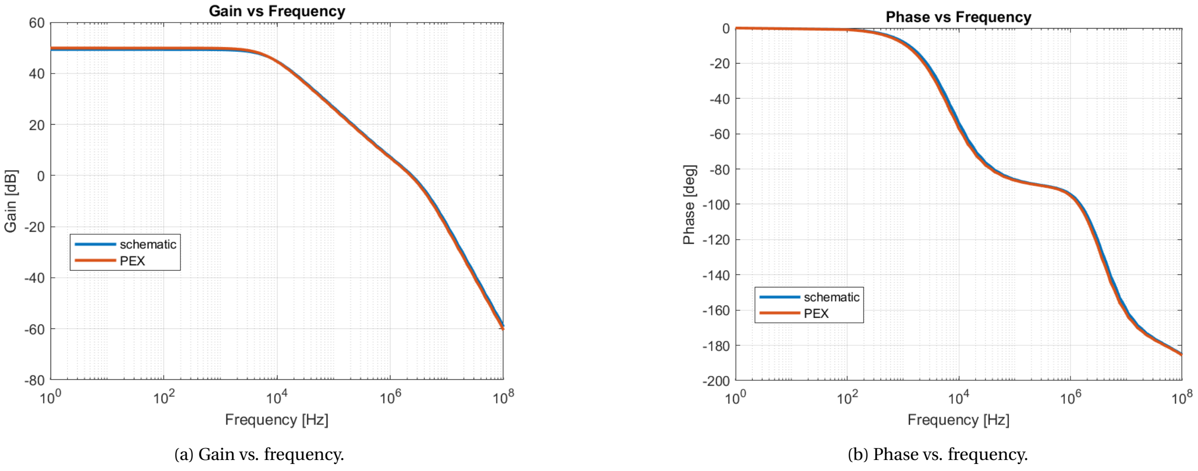


Figure 4.23: Opamp performance across corners.

Figure 4.24 presents the AC response of the opamp in both schematic level- and extracted simulations. Although bandwidth and phase margin degrade slightly due to parasitic capacitance, a phase margin above 60 degrees can still be achieved with sufficient speed.



(a) Gain vs. frequency.

(b) Phase vs. frequency.

Figure 4.24: AC response of the opamp.

To make sure the opamp is robust to process variation and mismatch, 200 runs of Monte Carlo simulation are done. Although the performance differs slightly, it does not affect the overall accuracy of the voltage reference as these nonidealities have already been taken into account during the error budgeting. The total current consumption of the opamp is 17.6 µA at 27 °C. The main opamp consumes 12.2 µA, the auxiliary opamp consumes 1.5 µA and the biasing network consumes 4 µA.

4.5. Clock Generator

The integrator requires a non-overlapping clock to ensure the charge stored on the capacitor will not leak away. The non-overlapping clock generator in this design is formed by NOR gates and inverters, as shown in figure 4.25. Compared with the conventional non-overlapping clock generator, additional logic gates are embedded in the circuit to allow two-channel operation. For example, when the integrator is in hold mode, S_4 is always closed and S_3 is always open (figure 4.26) such that the voltage stored at C_{int} can be held. The lower gate voltage is applied to S_4 to reduce the gate leakage. To allow more freedom to tune down this value, inverters (inv1-inv4) are implemented with low-threshold voltage devices. C_1 and C_2 are added to ensure there is sufficient non-overlapping time. With the capacitance value of 600 fF, 12 ns of nonoverlapping time is achieved.

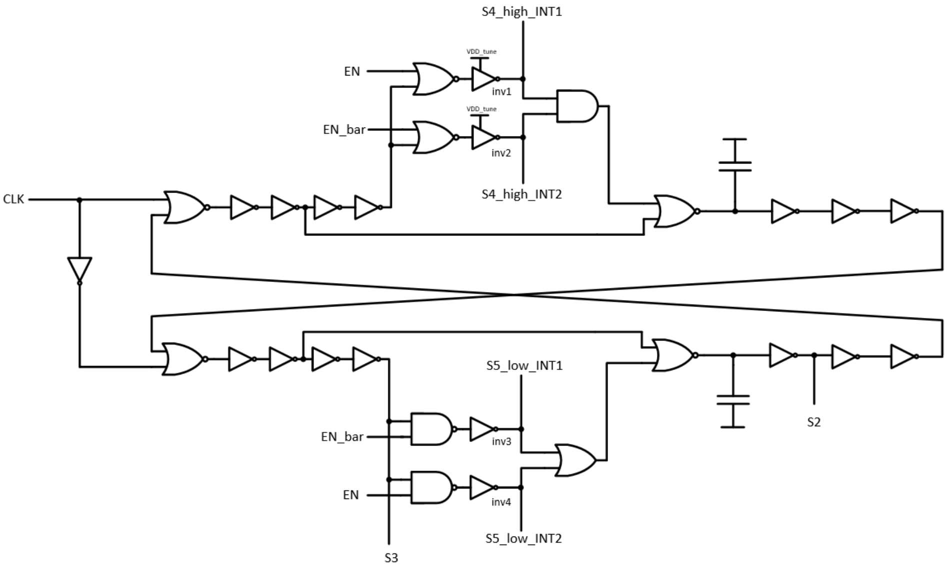


Figure 4.25: Non-overapping clock generator.

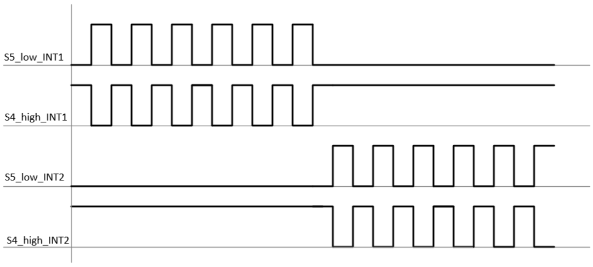


Figure 4.26: Timing for two-channel operation. $INT_{1,2}$ refers to $channel_{1,2}$.

All dynamic control signals for DEM, chopping, and averaging are generated on-chip. The generation of these signals is based on ring-counters. Figure 4.27 illustrates the generation of the chopping signal. Before the voltage reference starts working, the flip-flops will be reset to logic 0. On every rising edge of the clock signal, the logic 0 will be shifted to the next flip-flop. Consequently, the bitpattern will circulate through the ring. For each chopping phase, there are 16 settings for DEM. Therefore, there are in total 16 flip-flops. A negative-edge-triggered flip-flop is added at the end of the chain as a dummy. By doing so, race conditions can be prevented. The control signals for the two-channel operation, changing V_{cm}, and reset are also generated in the same way. Note that this is essentially a shift register connected in a ring-structure. The clock will stop when the integrator is in the reset period.

Similar to the generation of the control signals for the chopping, the control signals for DEM are also generated by a ring counter. Two different types of flip-flops are used for this purpose as shown in figure 4.28.

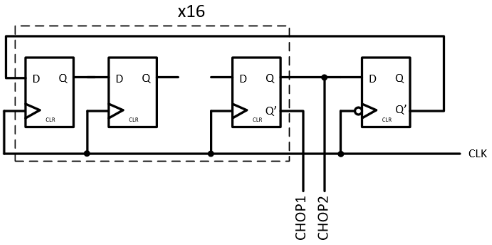


Figure 4.27: Clock generation for chopping signals.

In this figure, $DEM_{L,M,R}$ refers to the control signals for the M_{s1-3} switches in figure 4.3, respectively and S2 refers to the control signal for switch S_2 in figure 4.11a. Before starting the DEM procedure, fixed patterns will be hard-coded to the ring counter by setting some flip-flops to 1 and some to 0. During the DEM procedure, this bitpattern is circulated through the ring. Between two chopping phases, the current ratio between I_{left} and I_{right} (figure 4.3) has to be interchanged. Therefore, multiplexers are used to interchange the DEM signals generated by the ring counter. During the integrating phase of the SC-integrator, M_{s3} is opened to cut off the current at the output branch. This is implemented by an OR gate together with the signal S2.

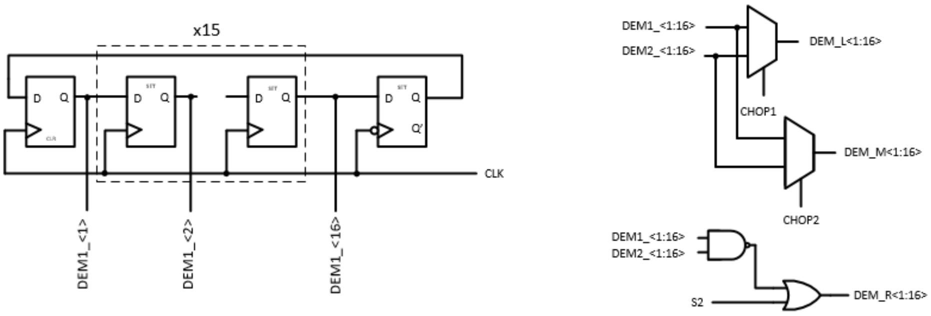


Figure 4.28: Clock generation for DEM signals.

4.6. Full Chip Overview

4.6.1. Layout

This design is implemented in a 40-nm LP bulk CMOS technology. The full chip dimensions are 1.085×1.085 mm^2 and the layout is shown in figure 4.29. The designed voltage reference is shown in the same figure. The total layout area is 167×562 μm^2, including two versions of the reference generator (as explained below), the averaging circuit, the clock generator, and auxiliary circuits. Each reference generator can be enabled and be selected to the averaging circuit separately.

Several considerations have been taken into account when designing the layout. In general, mismatch is important for almost all analog blocks, as mismatch degrades the accuracy of the reference voltage. Accordingly, the matching-sensitive transistors are implemented based on unit transistors with a fixed width and length, such that scaling factors can be generated based on a ratio in the number of unit elements. Furthermore, matching-sensitive transistors all have the same orientation and have dummy transistors on the two sides of the arrays. Since offset cancellation techniques are only applied to the reference generator, the input pair and current sources of the opamp are laid out in a common-centroid way to minimize mismatch caused by gradients over the chip. Moreover, it ensures that the environment seen by all transistors is as equal as possible.

To avoid coupling between two channels causing hold-mode feedthrough, the routing inside the averaging circuit is aimed to have a low coupling capacitance between each other. The width of the routing is kept close to the minimum value to minimize parasitic capacitance. Furthermore, the critical nodes are shielded.

Next to the reference generator presented in section 4.2.1, another reference generator is implemented on-chip. This additional reference generator is exactly the same as the one presented in [6] in terms of architecture and sizing. Adding this additional reference generator allows for investigating whether the kink observed in the PTAT voltage (3.3) is depending on the channel length of the core devices. To ensure a sufficient amount of samples for the statistical measurements, three samples are placed on the chip. The output of each reference can be multiplexed to a common output that can be measured. The switches used for measurement purposes are implemented with 2.5 V thick-oxide devices to ensure sufficiently low on-resistance and to reduce gate leakage. The remaining area on the chip is filled with de-coupling capacitors to minimize supply noise.

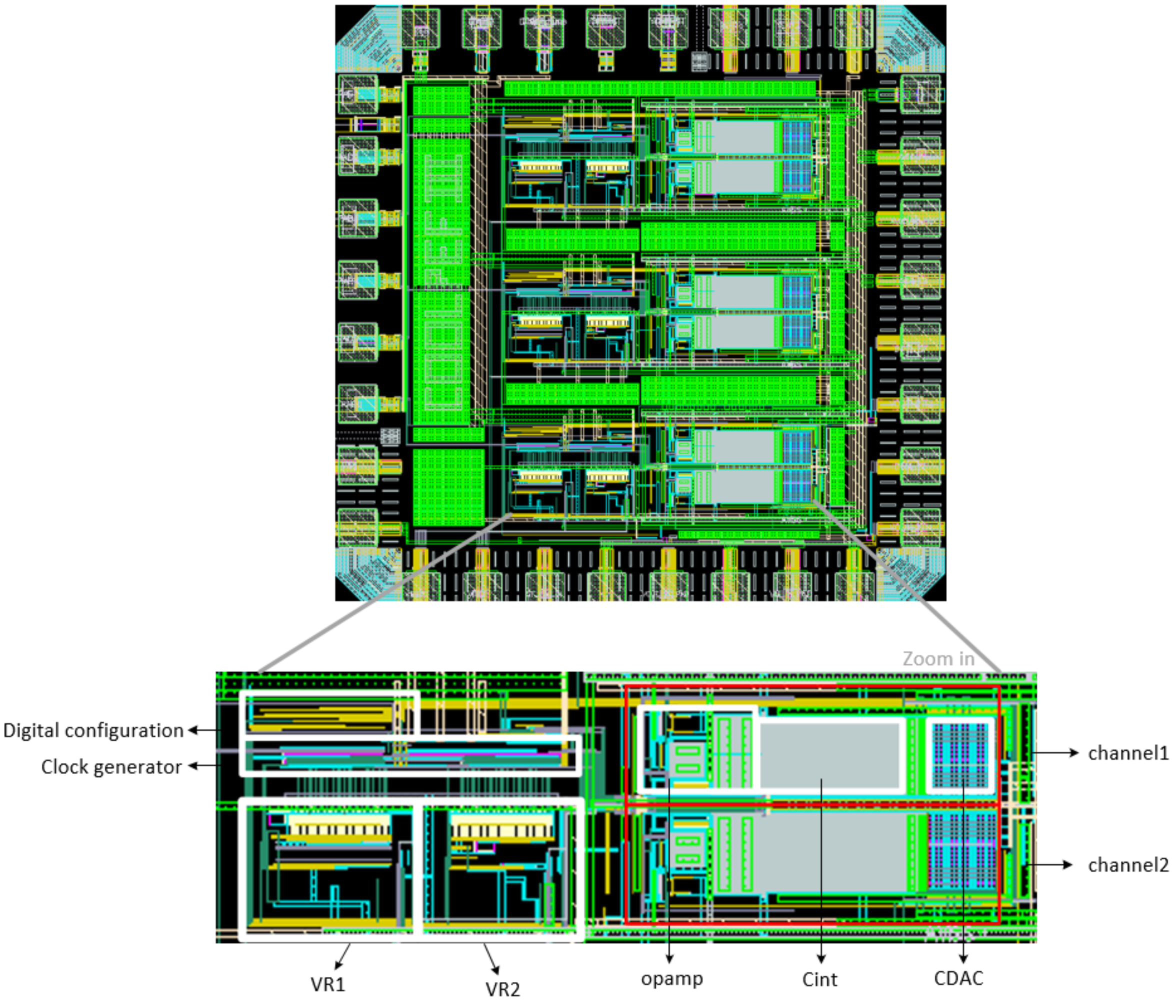


Figure 4.29: Full-chip layout image.

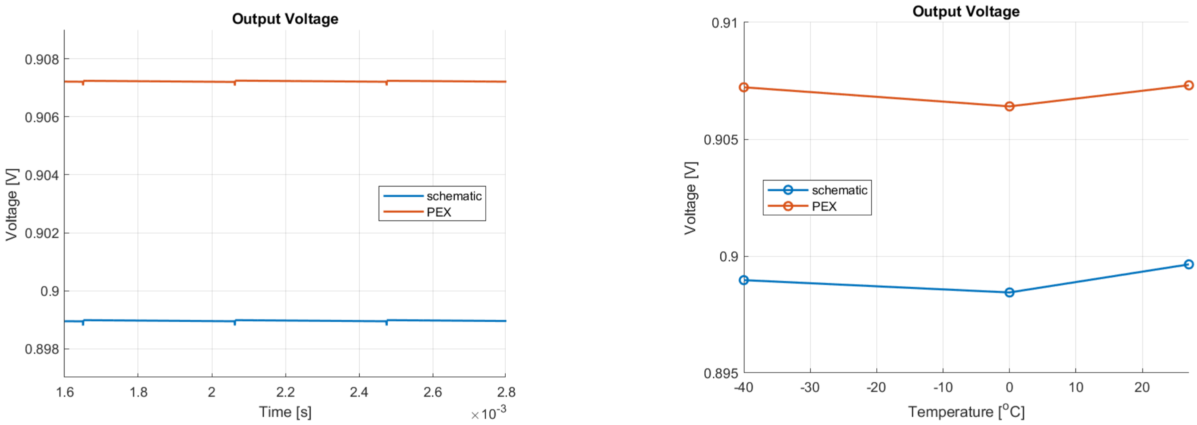
5

Performance of the Voltage Reference

This chapter presents the simulation results of the full system as shown in figure 4.1. The main objective of this chapter is to quantify the performance of the designed voltage reference in terms of temperature coefficient and inaccuracy (3σ). Furthermore, the simulation results regarding line regulation, noise, and power consumption will be presented. Afterwards, a comparison with the target specifications and a comparison with state-of-the-art is presented. In this chapter, the output of the reference generator will be referred to as V_{ref} and the final output of the designed voltage reference will be referred to as V_{out}.

5.1. Typical V_{out}

This section presents the typical output voltage V_{out} of the final design for both schematic- and parasitic extracted (PEX) simulations. The results are obtained by simulating the full system including all the analog-, digital- and auxiliary blocks. The netlist used for simulation is extracted under with the C+CC setting. Figure 5.1a shows V_{out} at $-40\,°\text{C}$ in the time domain while the nominal V_{out} at different temperatures is shown in figure 5.1b. V_{out} is larger after parasitic extraction, which can be attributed to the layout-induced threshold voltage shift in V_{ref} as mentioned in section 4.2.4.



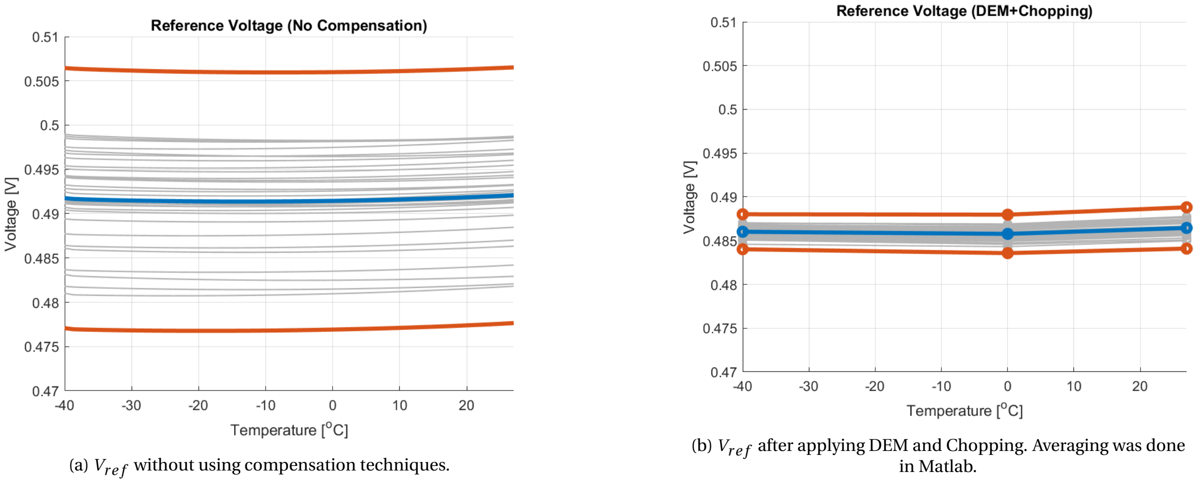
(a) V_{out} vs. time at $-40\,°\text{C}$.

(b) V_{out} vs temperature.

Figure 5.1: V_{out} from both schematic and parasitic extracted (PEX) simulations.

5.2. TC and inaccuracy (3σ)

Chapter 3 pointed out that mismatch is a bottleneck in achieving a high-accuracy voltage reference. Based on the measurement data, when applying DEM, chopping, and a scaling trim, it is possible to reach a 3σ of 1.2 %, compared to 5.1 % in the case of no compensation. The simulation results presented in this section are at the schematic level and with the nominal settings for the trimming networks, i.e., the PTAT- and scaling trimming code are both set to the mid-code. The results are obtained from 32 runs of Monte Carlo simulations at TT corner. Due to the significant amount of simulation time, simulations were performed only at three temperature points (−40 °C, 0 °C, and 27 °C).



(a) V_{ref} without using compensation techniques.

(b) V_{ref} after applying DEM and Chopping. Averaging was done in Matlab.

Figure 5.2: V_{ref} with and without applying compensation techniques.

Figure 5.2 shows the simulated V_{ref} with and without applying DEM and chopping. The blue lines show the average V_{ref} while the red lines represent the ±3σ. Figure 5.2a represents the same simulation as figure 4.7a, but with 32 samples for a fair comparison in 3σ with 5.2b. To give a clear idea of how offset compensation can improve performance, averaging was done in Matlab to avoid the nonidealities in the averaging circuit degrading the improvement. The simulation results were first obtained from the transient simulation, after that, each DEM and chopping phase was sampled manually in Matlab and the average V_{ref} was computed based on all the phases. The TC that the reference generator can achieve over the range of −40 °C to 27 °C is 554 ppm/K before and 106 ppm/K after compensation, respectively. The 3σ reduces from 3 % to 0.5 %. The remaining errors are mainly from the threshold voltage spread of the output transistor (M_6 in figure 4.2) and resistor mismatch and spread.

Figure 5.3 shows the transient response of the full system at −40 °C. Different curves represent different samples. Square-wave-like ripples in V_{out} are caused by the mismatch between the 2 channels (figure 4.11a). By using different trimming codes for two channels when performing the scaling trim, the ripples can be minimized. However, since transient simulations with fine accuracy settings take significant amounts of time, a scaling trim has not been extensively simulated.

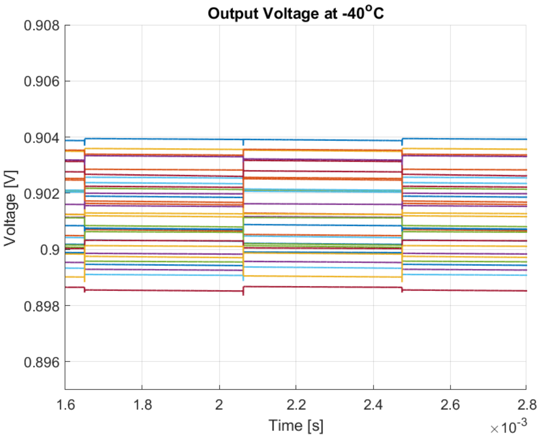
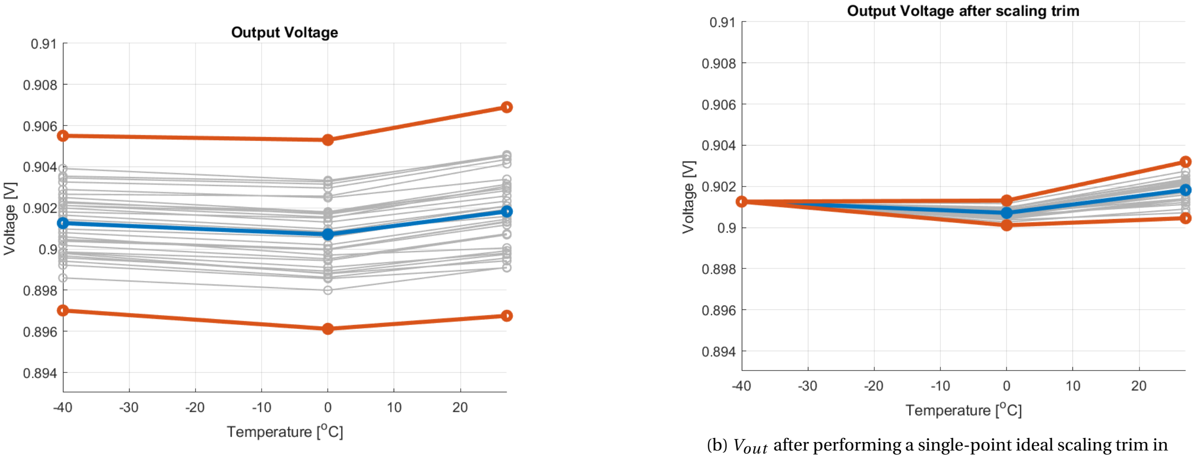


Figure 5.3: V_{out} vs. time at −40 °C.

Figure 5.4a shows V_{out} of the designed voltage reference over temperature. Without performing trimming, a TC of 109 ppm/K and 3σ of 0.56 % can be achieved. However, compared with the ideal averaging, the actual performance that this design achieves shows a slight degradation of 3 ppm/K. This is attributed to the non-idealities in the averaging circuit, such as leakage current, capacitor mismatch, and nonidealities from the opamp. Figure 5.4b shows the result after performing a single-point scaling trim in Matlab at −40 °C. Although performing trimming at 0 °C results in better performance at room temperature simulation, performing trimming at −40 °C is closer to the actual measurement. Compared with the result shown in figure 5.4a, it is clear that the offset errors, such as threshold voltage spread, can be effectively removed by performing the scaling trim. A TC of 40.9 ppm/K and 3σ of 0.15 % are achieved. Note that the actual CDAC is designed to have a resolution of 0.5 mV. The actual scaling trim that will be performed in the measurements cannot trim out the offset error as effectively as in Matlab. Since 0.5 mV only translates to 2 ppm/K in TC, the resolution of the scaling trim will not limit the overall performance.



(a) V_{out} obtained from the ideal integrator and actual design.

(b) V_{out} after performing a single-point ideal scaling trim in Matlab.

Figure 5.4: V_{out} before and after trimming.

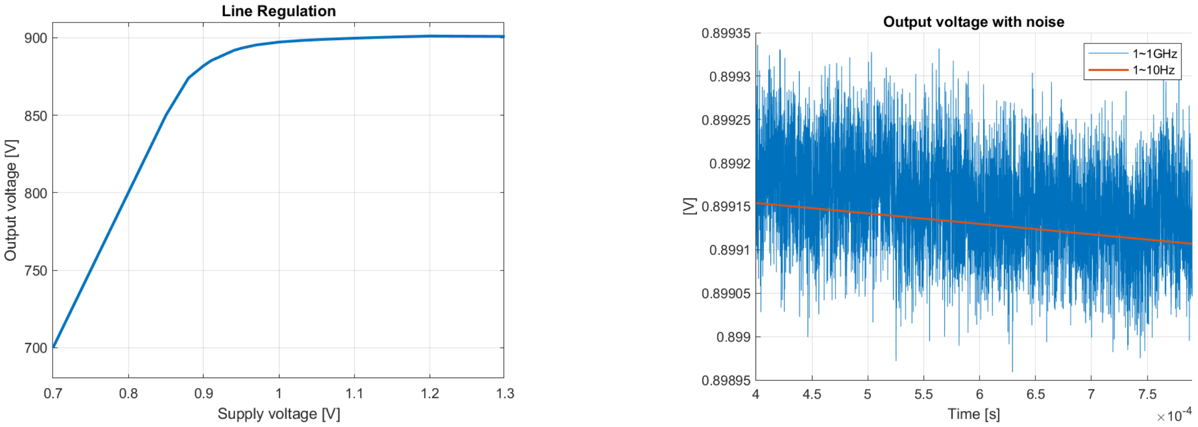
Table 5.1 summarizes the TC and 3σ before and after applying the offset cancellation techniques. Based on room temperature simulations, a 20 times improvement in 3σ and a 13.5 times improvement in TC have been reached after performing a single-point scaling trim at −40 °C.

Table 5.1: Achieved accuracy with and without applying offset cancellation techniques.

	TC [ppm/K]	3σ [%]	figure
Intrinsic accuracy (without compensation)	553.7	3	5.2a
Ideal averaging	105.8	0.49	5.2b
Achieved accuracy (without trimming)	109	0.56	5.4a
Achieved accuracy (after performing ideal scaling trim)	40.9	0.15	5.4b

5.3. Line Regulation

Figure 5.5a shows the line regulation simulation for the design. All the power supplies, including analog-, digital- and auxiliary- blocks were swept at the same time from 0.7 V to 1.3 V. For a 0.5 % error in V_{out}, the minimum required supply is 0.96 V, given that the nominal supply voltage of 1.1 V and the nominal V_{out} of 0.9 V. The simulated line regulation is 15.65 mV/V. The degradation when supply drops can be attributed to the following reasons. Firstly, when the supply goes down, the reference generator cannot generate the desired V_{ref} due to the limited output impedance of the output branch. Secondly, lower supply reduced the output swing of the opamp. Integration cannot reach the desired accuracy if the DC gain of the opamp is insufficient, therefore introducing errors.



(a) Line regulation.

(b) Transient with noise.

Figure 5.5: Noise and line regulation simulations.

5.4. Noise

This section illustrates how the noise, mismatch, and leakage limit the noise performance of the design. Figure 5.5b shows the results of a transient simulation with noise at −40 °C for one holding period (figure 4.16). The result is obtained by simulating the reference generator and one-channel only, with all the clocking signals being ideal. The blue curve shows the noise integrated from 1 Hz to 1 GHz while the orange curve only considers the low-frequency noise from 1 Hz to 10 Hz. The peak-to-peak value of the noise in the 1 Hz to 1 GHz bandwidth is approximately 373 µV. Thanks to the temperature dependence of the thermal noise, the high-frequency noise is expected to reduce 10x at 4 K compared with room temperature [48]. On the other hand, the voltage drop on V_{out} caused by gate leakage presented itself as low-frequency noise and it is approximately 47 µV (orange curve in the figure). Gate leakage can be assumed to be temperature-independent, and it is therefore not expected to significantly change at 4 K. Although it cannot be seen from figure 5.5b, the mismatch between the two channels is the main limiting factor that degrades the noise performance as illustrated in figure 5.3. Since two channels are interchanged at a low frequency (in the range of 300 Hz to 2.4 kHz) the ripple caused by this switching effect is harder to be filtered out compared with high-frequency noise.

5.5. Power Consumption

Table 5.2 shows the power breakdown for the main blocks at both 27 °C (simulated) and 4 K (expected). At 27 °C, the opamp consumes the largest amount of power, followed by the reference generator. The overall power consumption is 57 µW, including one type of reference generator (v1), two opamps, a clock generator, and the auxiliary circuits. Due to the PTAT biasing in the reference generator and opamp, the power consumption of these blocks is expected to reduce roughly five times at 4 K based on the measurement data. Therefore, the power consumption is expected to be 17.7 µW at 4 K. The power consumption of the design at 4 K is limited by the resistive dividers (for the generation of V_{cm} and V_{ls} as mentioned in section 4.3.7 and section 4.4.2) in the auxiliary block, which consume 43 % of the total power. This can be reduced further in the future design by making a trade-off with the area.

Table 5.2: Power breakdown of the main blocks at 27 °C (simulated) and 4 K (expected).

	Power consumption [W]	
	27 °C (simulation)	4 K (expected)
Reference generator (v1)	10.5μ	2.1μ
Reference generator (v2)	11.9μ	2.38μ
Opamp	19.36μ	3.87μ
Clock generator	$231n$	$231n$
Auxiliary circuit	7.6μ	7.6μ

5.6. Comparison with the State-of-the-Art

Table 5.3 summarizes the primary achieved/expected performance and gives a comparison with the target specifications and the state-of-the-art performance. Achieved performance is the result based on room temperature simulation while the expected performance is derived based on the measurement data (figure 3.16). Extended work of [6] has been presented in chapter 3. The expected performance of the proposed design outperforms other cryogenic voltage references in terms of TC and inaccuracy (3σ). However, it still does not reach the target TC. The TC obtained from room temperature simulation is mainly limited by the higher-order errors present in the CTAT voltage as mentioned in section 3.3.2. Since the higher-order nonlinearities in CTAT voltage are not the bottleneck for designing a cryogenic voltage reference, no further technique has been applied to compensate for this. The expected TC over the full temperature range from 4 K to 300 K is mainly limited by:

- **Output transistor spread:** The spread in the output transistor (M_6 in figure 4.2) will translate to mismatch with respect to other samples. Although the effect of the threshold voltage spread can be effectively minimized by performing a scaling trim, the threshold voltage mismatch also contains a temperature-dependent part as discussed in section 3.4.1. The temperature-dependent part of the threshold voltage mismatch cannot be removed by the scaling trim. Consequently, this is a bottleneck for further improving the TC and 3σ.

- **Second-order effect:** The idea of DEM is to average the mismatch errors instead of fully canceling them [16]. As a result, there will be systematic nonlinearities left. Besides, there will be residual errors caused by the second-order effect in the DEM procedure.

- **Resistor mismatch:** Since the scaling trim is more effective than the PTAT trim (table 3.1), the PTAT trim will only be used as a bath-trim in this design to correct the systematic PTAT error. However, not all the samples have exactly the same PTAT errors, which limits the effect of the PTAT trim. It is possible to further remove these errors by performing PTAT trim on each sample. However, it will require performing a 2-point trimming, which is expensive in terms of time and cost.

Future recommendations to allow further improving the performance will be given in section 6.2.

Table 5.3: Comparison table with target specifications and the state-of-the-art.

	Achieved (RT sim)	**Expected (full-T)**	**Target specs.**	ESSCIRC 2019 [6] (extended work)	SSCL 2020 [31]	JSSC 2021 [20]
Technology	40-nm COMS			40-nm CMOS	28-nm FDSOI	0.18-um CMOS
Supply [V]	1.1			1.1	1.2	1.8
Temp. range [K]	233 to 300	4 to 300		4 to 300	4 to 300	233 to 398
Vref [V]	0.899	0.9	0.9	0.5	0.66	1.14
TC [ppm/K]	40.9	60	10	257.2	1214	5.5
Trimming points	1	1	1	1	-	1
Inaccuracy (3σ) [%]	0.15	1	1	3.8	NA	0.14

6
Conclusion

6.1. Main conclusions

This thesis presents a voltage reference that can work over a wide temperature range from 4 K to 300 K. The proposed design aims to achieve comparable performance with the state-of-the-art voltage references working at room temperature in terms of TC and inaccuracy (3σ). To overcome the challenge of the lack of reliable device models, this thesis first provides an understanding of the temperature dependence of the PTAT- and CTAT voltage at cryogenic temperature. Furthermore, the effects of process variations on the performance of voltage references are investigated. Based on the analysis, the voltage reference presented in this thesis is designed according to the devices' cryogenic behaviour to minimize the nonlinearities present in the PTAT- and CTAT voltage. Several compensation techniques, such as dynamic element matching, chopping, and trimming are implemented to reduce the effects of process variations.

This design uses a switched-capacitor circuit to effectively remove the ripples introduced by DEM and chopping. Two SC-integrators are working in a ping-pong mode to obtain a continuous output. To cope with the uncertainties at cryogenic temperatures, the input sampling switch used in conventional CDS-integrator architectures has been removed, and the conventional non-overlapping timing has been adjusted to allow for this modification. The opamp used in the integrator is a gain-enhanced active-current mirror OTA. It features a large output swing and is power efficient. Besides, it is able to handle the low common-mode input voltage thanks to the active current mirrors.

The design has been simulated from −40 °C to 27 °C, where it achieves a TC of 40.9 ppm/K and inaccuracy (3σ) of 0.15 % after performing a single-point scaling trimming. Compared to the case where no compensation is applied, 3σ shows a promising improvement of almost 20 times from the simulating range. With the nominal output voltage of 0.9 V, the minimum required supply voltage can be as low 0.96 V while the error in V_{out} is still less than 0.5 %. The power consumption is 57 µW at 27 °C and it is expected to reduce 5x at cryogenic temperatures. The design has been taped out in a 40-nm bulk CMOS process. To evaluate the performance from 300 K down to 4 K, measurements will be performed after receiving the chips.

6.2. Future work

This section aims to provide recommendations for future work based on the simulation results presented in chapter 4 and chapter 5. These recommendations can be summarized as follows:

- The design has been tape-out in August. In order to verify the functionality and performance of the chip, the measurements will be done afterwards. The designed voltage reference will be measured from room temperature down to 4 K. Besides measuring the overall performance, two versions of the reference generator and the integrator (figure 4.29) will be characterized.
- Chapter 3 investigated the temperature dependence of the PTAT- and CTAT voltage over a wide temperature range. However, some of the physical effects behind this temperature dependence are still ongoing research. Further study regarding device physics can be done to gain more insights into the temperature dependence of the devices.

- It is expected that the large curvature observed in the reference voltage (figure 3.2) at cryogenic temperature can be fixed once there is more understanding regarding the temperature dependence of the device. After correcting this limiting error,a curvature correction technique can be applied to reduce the higher-order nonlinearities in both the PTAT- and CTAT voltage.

- Due to the low chopping frequency, the capacitors that are used in this design are comparatively large to minimize the effect of leakage. Large capacitance is costly in terms of area. The chopping frequency is currently limited by the reference generator due to the large transistors. The reference generator can be resized to allow for using a higher chopping frequency.

- Two-channel operation comes with some inherent drawbacks. For example, higher power consumption, larger area, mismatch between two channels, spikes introduced by the switching, etc. Another averaging approach can be used in the future. Moreover, if the voltage reference can be integrated with the following circuit blocks, it is possible to perform the averaging in those blocks.

- The resolution of the designed voltage reference is currently limited by the mismatch between the two integrator channels. To minimize the effect of mismatch, the resolution of the scaling trim network (CDAC) can be increased. In addition, the error budgeting for the averaging circuit in this project is decided based on the targeted 3σ instead of the resolution. The resolution degradation caused by mismatch can be alleviated by taking the noise performance into account when doing the error budgeting.

- The designed voltage reference aims to generate a buffered output. However, due to the low current level, the output impedance of this design is not sufficiently low to drive low-impedance nodes. Consequently, the driving strength of the voltage reference can be improved in the future.

- The critical nodes in the design have been shielded to avoid capacitive coupling. However, there might still be coupling from the top or bottom plate of the integration capacitor. To minimize the coupling, the integration capacitor can be fully shielded.

- The analysis in chapter 3.4.5 indicates that there will still be remaining errors even after performing the offset compensation techniques, and assuming that it is possible to design a perfect curvature correction circuit. The remaining mismatch is a bottleneck for achieving a low TC at cryogenic temperatures. It is mainly caused by second-order effects, such as resistor mismatch, and the mismatch in the output transistor with respect to other samples. Resistor mismatch can be minimized by making a trade-off with the current level while the second-order effects can be alleviated by careful layout.

- Finally, once the high-accuracy, low TC, voltage reference is available, it would be good to integrate the design with other electronic blocks.

A

Error Sources Transfer Function

The architecture used for deriving the error sources transfer is shown in figure 3.1. With M_1, M_2 being the core transistors, M_6 being the output transistor, M_{3-5} being current sources, having the ratio of 1:p:m, and R_1, R_2 are PTAT resistor (R_{ptat}) and output resistor (R_{out}) separately.

V_{ref} can be expressed as follows,

$$V_{ref} = V_{gs} + \gamma \Delta V_{gs}, \tag{A.1}$$

where γ is the scaling factor

$$\gamma = m \times \frac{R_{out}}{R_{ptat}}. \tag{A.2}$$

The effect of error sources on V_{ref} is obtained by multiplying the error of a specific instance with respective sensitivities. The sensitivity functions with respect to each major error source: PTAT voltage, CTAT voltage, and scaling factor, are summarized in equations A.3 to A.5. Derivations are given in the following subsections.

$$S_{V_{gs}}^{V_{ref}} = \frac{\partial V_{ref}}{\partial V_{gs}} \approx 1 \tag{A.3}$$

$$S_{\Delta V_{gs}}^{V_{ref}} = \frac{\partial V_{ref}}{\partial \Delta V_{gs}} \approx \gamma + \frac{1}{ln(p)} \tag{A.4}$$

$$S_{\gamma}^{V_{ref}} = \frac{\partial V_{ref}}{\partial \gamma} \approx n\frac{kT}{q}\left(ln(p) + \frac{1}{\gamma}\right) \tag{A.5}$$

A.1. PTAT Voltage (ΔV_{gs})

Same as ΔV_{gs}, the error in ΔV_{gs} is scaled by the scaling factor γ and then added to the reference voltage. In addition, a change in PTAT current affects the bias current of the output transistor M_6. Therefore, the sensitivity function A.4 consists of two terms:

- **γ (Error in ΔV_{gs})**
  Directly differentiate equation A.1 gives the following sensitivity function

  $$S_{\Delta V_{gs}|core}^{V_{ref}} = \frac{\partial V_{ref}}{\partial \Delta V_{gs}} \approx \gamma, \tag{A.6}$$

  which indicates that the error in ΔV_{gs} has an effect on V_{ref} depending on the scaling factor.

- **1/ln(p) (Error in I_{M6})**
  Error in ΔV_{gs} ($\delta \Delta V_{gs}$) appears in the PTAT current, and consequently, the biasing current of the output transistor M_6 also contains this error. With the ideal current at the output branch, I_0 and this error term δI_0, the output current I_d can be expressed as

  $$I_d = I_0 + \delta I_0. \tag{A.7}$$

This error term δI_0 is then converted into the error in V_{ref} by the output resistor R_2,

$$\delta V_{ref} = \delta I_0 \times R_{out} = \gamma \times \delta\Delta V_{gs}, \tag{A.8}$$

where δV_{ref} is the error on V_{ref} and δI_0 cab be approximated as

$$\delta I_0 = \frac{\delta\Delta V_{gs}}{R_{ptat}} \times m. \tag{A.9}$$

Given that M_6 is in weak inversion, I_d is given by

$$I_d = \frac{\Delta V_{gs} + \delta\Delta V_{gs}}{R_{ptat}} m = \frac{nV_T ln(p)}{R_{ptat}} m = \frac{W}{L}\mu_n C_{ox} exp\left(\frac{V_{gs} - V_{th}}{nV_T}\right) \tag{A.10}$$

V_{gs6} can be obtained by re-arranging equation A.10,

$$V_{gs6} = nV_T ln(I_d) - nV_T ln(C) + V_{th}. \tag{A.11}$$

From the above equations, it is obvious that any change in I_d will cause a change in V_{gs}. The sensitivity can be derived as follows,

$$\frac{\partial V_{gs}}{\partial I_d} = \frac{nV_T}{I_d} = \frac{nV_T}{\frac{nV_T ln(p) m}{R_{ptat}}} = \frac{R_{ptat}}{m ln(p)} \tag{A.12}$$

Therefore, the change in V_{gs} with respect to the change in I_d can be calculated as

$$\delta V_{gs} = \frac{\partial V_{gs}}{\partial I_d} \times \delta I_0 = \frac{R_{ptat}}{m ln(p)} \frac{\delta\Delta V_{gs}}{R_{ptat}} m = \frac{\delta\Delta V_{gs}}{ln(p)}, \tag{A.13}$$

$$\frac{\delta V_{gs}}{\delta\Delta V_{gs}} = \frac{1}{ln(p)}. \tag{A.14}$$

A.2. Scaling Factor γ

ΔV_{gs} will translate to PTAT voltage after scaling by γ. Besides, due to a change in the biasing current of M_6, V_{gs} will also change. The sensitivity function A.5 consists of the following two terms:

- $nV_T ln(p)$ **(Error in ΔV_{gs})**
  Ideal reference voltage V_{ref} is given in equation A.1. Therefore, to the first order, any change in the scaling factor γ causes an error in the reference voltage through the sensitivity given below,

$$\frac{\partial V_{ref}}{\partial \gamma} \approx \Delta V_{gs} = n\frac{kT}{q} ln(p). \tag{A.15}$$

- V_T/γ **(Error in I_{M6})**
  Since a change in the scaling factor alters the current at the output branch, V_{gs} is affected.

$$I_d = m \times \frac{\Delta V_{gs}}{R_{ptat}} = \gamma \frac{\Delta V_{gs}}{R_{out}} \tag{A.16}$$

The output current I_d, consists of the error term δI_d can be expressed as A.7, where

$$\delta I_0 = \delta\gamma \times \frac{\Delta V_{gs}}{R_{out}}. \tag{A.17}$$

Through the sensitivity function derived in equation A.12, the effect of δI_d on δV_{gs} due to the error in scaling factor can be expressed as

$$\delta V_{gs} = \frac{\partial V_{gs}}{\partial I_d} \times \delta I_0 = \frac{R_{ptat}}{m ln(p)} \times \left(\delta\gamma \times \frac{\Delta V_{gs}}{R_{out}}\right). \tag{A.18}$$

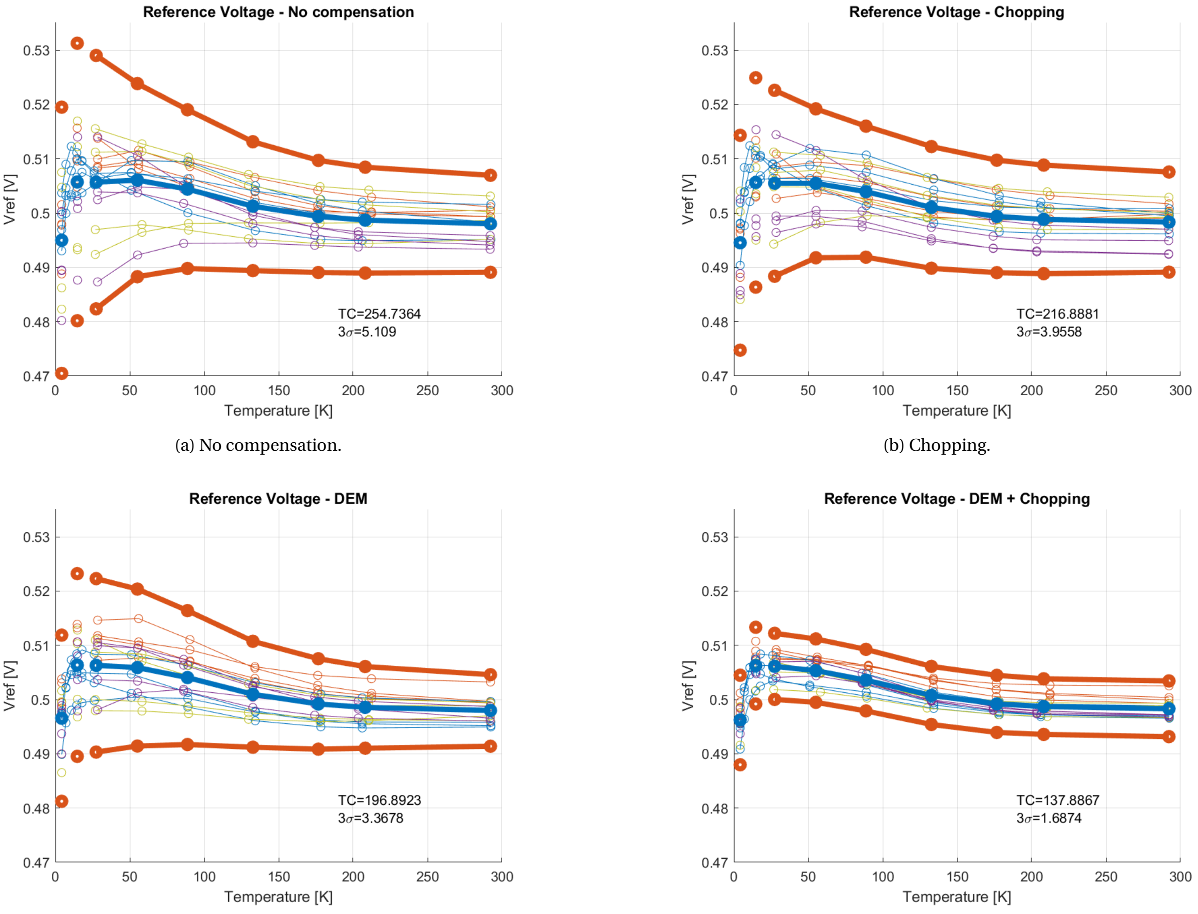
Re-arranging the above equation yields

$$\frac{\delta V_{gs}}{\delta\gamma} = n\frac{kT}{q}\frac{1}{\gamma}. \tag{A.19}$$

B

Effect of Compensation Techniques on TC and 3σ

This appendix presents the effect of compensation techniques on TC and 3σ. The measurements shown in this appendix are from 16 samples of the NMOS-base reference with switches that allow for DEM on both the current sources and core transistors. Different combinations of the compensation technique as introduced in section 3.4 are applied. The PTAT- and scaling trim are performed in Matlab based on the interpolation data from the measurements. The TC and 3σ labeled on the figures have the unit of ppm/K and %, respectively.



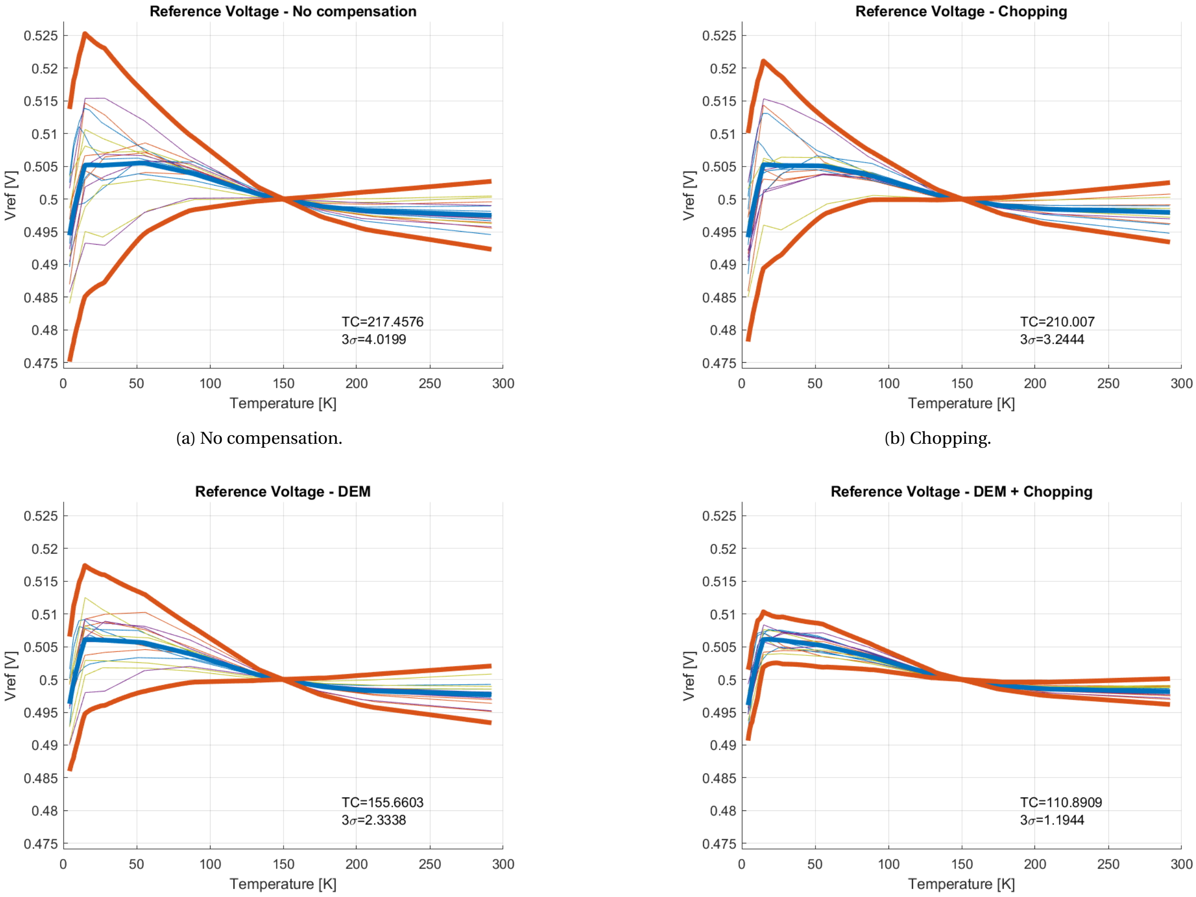
(a) No compensation.

(b) Chopping.

(c) DEM.

(d) DEM + chopping.

Figure B.1: Reference voltage with and without applying DEM and chopping (untrimmed).



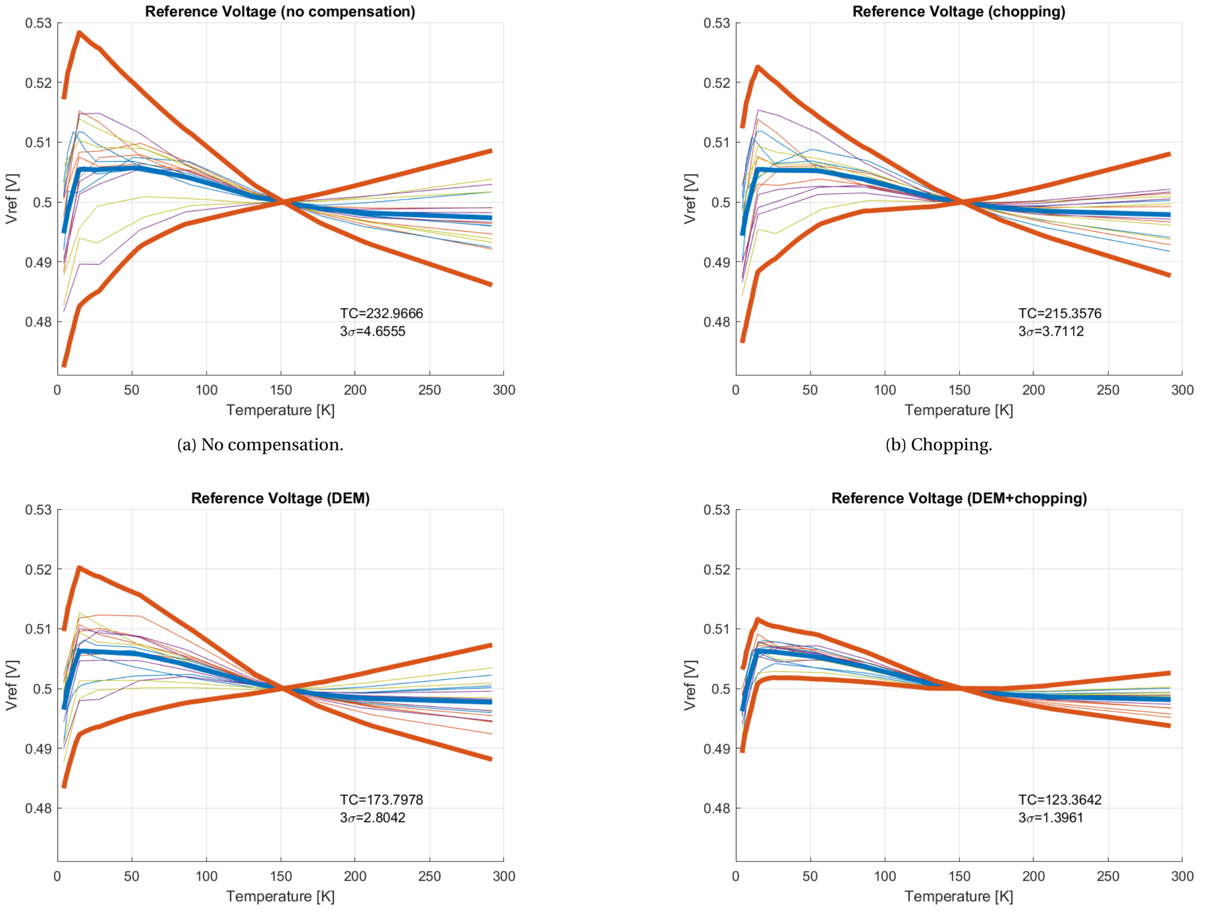
(a) No compensation.

(b) Chopping.

(c) DEM.

(d) DEM + chopping.

Figure B.2: Reference voltage with and without applying DEM and chopping. With a single-point scaling trim at 150 K.



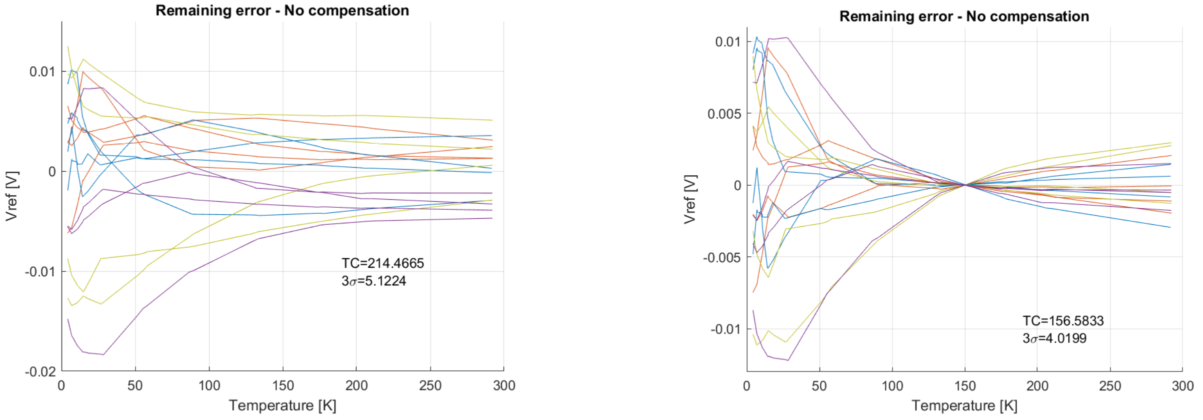
(a) No compensation.

(b) Chopping.

(c) DEM.

(d) DEM + chopping.

Figure B.3: Reference voltage with and without applying DEM and chopping. With a single-point PTAT trim at 150 K.



(a) Curvature correction (untrimmed).

(b) Curvature correction + single-point scaling trim at 150 K.

Figure B.4: Remaining errors after applying curvature correction.

Bibliography

[1] F. Jazaeri, A. Beckers, A. Tajalli, and J.-M. Sallese, "A Review on Quantum Computing: From Qubits to Front-end Electronics and Cryogenic MOSFET Physics," in *2019 MIXDES - 26th International Conference "Mixed Design of Integrated Circuits and Systems"*, 2019, pp. 15–25.

[2] L. Vandersypen and A. van Leeuwenhoek, "1.4 Quantum computing - the next challenge in circuit and system design," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 24–29.

[3] E. Charbon, F. Sebastiano, A. Vladimirescu, H. Homulle, S. Visser, L. Song, and R. M. Incandela, "Cryo-cmos for quantum computing," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 13.5.1–13.5.4.

[4] B. Patra, R. M. Incandela, J. P. G. van Dijk, H. A. R. Homulle, L. Song, M. Shahmohammadi, R. B. Staszewski, A. Vladimirescu, M. Babaie, F. Sebastiano, and E. Charbon, "Cryo-CMOS Circuits and Systems for Quantum Computing Applications," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, 2018.

[5] A. Hammoud, R. Patterson, S. Gerber, and M. Elbuluk, "Electronic components and circuits for extreme temperature environments," in *10th IEEE International Conference on Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003*, vol. 1, 2003, pp. 44–47 Vol.1.

[6] J. van Staveren, C. García Almudever, G. Scappucci, M. Veldhorst, M. Babaie, E. Charbon, and F. Sebastiano, "Voltage References for the Ultra-Wide Temperature Range from 4.2K to 300K in 40-nm CMOS," in *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*, 2019, pp. 37–40.

[7] P. Padalia, "Design of CMOS Voltage References for Ultra-Wide Temperature Ranges," Master's thesis, TU Delft, the Netherlands, 2019.

[8] W.-H. Tseng, W.-L. Lee, C.-Y. Huang, and P.-C. Chiu, "A 12-bit 104 MS/s SAR ADC in 28 nm CMOS for Digitally-Assisted Wireless Transmitters," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 10, pp. 2222–2231, 2016.

[9] M. J. Pelgrom, *Analog-to-Digital Conversion*, 3rd ed. Springer International Publishing Switzerland, 2017.

[10] K. Kuijk, "A precision reference voltage source," *IEEE Journal of Solid-State Circuits*, vol. 8, no. 3, pp. 222–226, 1973.

[11] D. Johns and K. Martin, *Analog Integrated Circuit Design*. Wiley India Pvt. Limited, 2008. [Online]. Available: https://books.google.com.tw/books?id=6-8j7ycydtcC

[12] H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, S. Atsumi, and K. Sakui, "A CMOS bandgap reference circuit with sub-1-V operation," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 5, pp. 670–674, May 1999.

[13] H. Homulle, L. Song, E. Charbon, and F. Sebastiano, "The Cryogenic Temperature Behavior of Bipolar, MOS, and DTMOS Transistors in Standard CMOS," *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 263–270, 2018.

[14] H. Homulle, F. Sebastiano, and E. Charbon, "Deep-Cryogenic Voltage References in 40-nm CMOS," *IEEE Solid-State Circuits Letters*, vol. 1, no. 5, pp. 110–113, 2018.

[15] L. Magnelli, F. Crupi, P. Corsonello, C. Pace, and G. Iannaccone, "A 2.6 nW, 0.45 V Temperature-Compensated Subthreshold CMOS Voltage Reference," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 2, pp. 465–474, 2011.

[16] M. A. P. Pertijs and J. H. Huijsing, *Precision Temperature Sensors in CMOS Technology (Analog Circuits and Signal Processing)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[17] G. Kiene, A. Catania, R. Overwater, P. Bruschi, E. Charbon, M. Babaie, and F. Sebastiano, "13.4 A 1GS/s 6-to-8b 0.5mW/Qubit Cryo-CMOS SAR ADC for Quantum Computing in 40nm CMOS," in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 64, 2021, pp. 214–216.

[18] L. Enthoven, J. van Staveren, J. Gong, M. Babaie, and F. Sebastiano, "A 3V 15b 157μW Cryo-CMOS DAC for Multiplexed Spin-Qubit Biasing," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2022, pp. 228–229.

[19] G. Ge, C. Zhang, G. Hoogzaad, and K. A. A. Makinwa, "A Single-Trim CMOS Bandgap Reference With a 3σ Inaccuracy of $\pm$0.15% From -40°C to 125°C," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp. 2693–2701, 2011.

[20] J.-H. Boo, K.-I. Cho, H.-J. Kim, J.-G. Lim, Y.-S. Kwak, S.-H. Lee, and G.-C. Ahn, "A Single-Trim Switched Capacitor CMOS Bandgap Reference With a 3σ Inaccuracy of +0.02%, -0.12% for Battery-Monitoring Applications," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 4, pp. 1197–1206, 2021.

[21] C.-Z. Shao, S.-C. Kuo, and Y.-T. Liao, "A 1.8nW, -73.5dB PSRR, 0.2ms Startup Time, CMOS Voltage Reference With Self-Biased Feedback and Capacitively Coupled Schemes," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 6, pp. 1795–1804, 2021.

[22] J. Jiang, W. Shu, and J. S. Chang, "A 5.6 ppm/°C Temperature Coefficient, 87-dB PSRR, Sub-1-V Voltage Reference in 65-nm CMOS Exploiting the Zero-Temperature-Coefficient Point," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 3, pp. 623–633, 2017.

[23] Y. Ji, J. Lee, B. Kim, H.-J. Park, and J.-Y. Sim, "18.8 A 192pW Hybrid Bandgap-Vth Reference with Process Dependence Compensated by a Dimension-Induced Side-Effect," in *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, 2019, pp. 308–310.

[24] P. A. 'T Hart, M. Babaie, E. Charbon, A. Vladimirescu, and F. Sebastiano, "PCharacterization and Modeling of Mismatch in Cryo-CMOS," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 263–273, 2020.

[25] A. Beckers, F. Jazaeri, and C. Enz, "Cryogenic MOSFET Threshold Voltage Model," in *ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC)*, 2019, pp. 94–97.

[26] A. Beckers, F. Jazaeri, A. Grill, S. Narasimhamoorthy, B. Parvais, and C. Enz, "Physical Model of Low-Temperature to Cryogenic Threshold Voltage in MOSFETs," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 780–788, 2020.

[27] P. A. T Hart, M. Babaie, E. Charbon, A. Vladimirescu, and F. Sebastiano, "Subthreshold Mismatch in Nanometer CMOS at Cryogenic Temperatures," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 797–806, 2020.

[28] A. Beckers, F. Jazaeri, and C. Enz, "Theoretical Limit of Low Temperature Subthreshold Swing in Field-Effect Transistors," *IEEE Electron Device Letters*, vol. 41, no. 2, pp. 276–279, 2020.

[29] R. M. Incandela, L. Song, H. Homulle, E. Charbon, A. Vladimirescu, and F. Sebastiano, "Characterization and Compact Modeling of Nanometer CMOS Transistors at Deep-Cryogenic Temperatures," *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 996–1006, 2018.

[30] Website. (2022) Cryogen-free dilution refrigerator measurement systems. [Online]. Available: https://bluefors.com/

[31] Y. Yang, K. Das, A. Moini, and D. J. Reilly, "A Cryo-CMOS Voltage Reference in 28-nm FDSOI," *IEEE Solid-State Circuits Letters*, vol. 3, pp. 186–189, 2020.

[32] F. Liu, Z. Deng, and Y. Liu, "Cryogenic Bandgap Reference Circuit With Compact Model Parameter Extraction of MOSFETs and BJTs for HPGe Detectors," *IEEE Transactions on Nuclear Science*, vol. 67, no. 10, pp. 2209–2216, 2020.

[33] A. Beckers, F. Jazaeri, and C. Enz, “Characterization and Modeling of 28-nm Bulk CMOS Technology Down to 4.2 K,” *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 1007–1018, 2018.

[34] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 1st ed. McGraw-Hill, 2001.

[35] A. Hastings and R. Hastings, *The Art of Analog Layout*. Prentice Hall, 2001. [Online]. Available: https://books.google.nl/books?id=v6WvQgAACAAJ

[36] M. Pelgrom, A. Duinmaijer, and A. Welbers, “PMatching properties of MOS transistors,” *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[37] J. van Staveren, “A Wide Temperature Range Voltage Reference for Quantum Computing Applications,” Master’s thesis, TU Delft, the Netherlands, 2018.

[38] B. Wu and S. Ay, “Low-Noise CMOS Bandgap Reference Generator Using Two-Level Chopping Technique,” in *2015 IEEE Workshop on Microelectronics and Electron Devices (WMED)*, 2015, pp. 1–4.

[39] W. Biederman, D. Yeager, E. Alon, and J. Rabaey, “A CMOS switched-capacitor fractional bandgap reference,” in *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, 2012, pp. 1–4.

[40] A. Shrivastava, K. Craig, N. E. Roberts, D. D. Wentzloff, and B. H. Calhoun, “5.4 A 32nW bandgap reference voltage operational from 0.5V supply for ultra-low power systems,” in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3.

[41] U. Chi-Wa, W.-L. Zeng, M.-K. Law, C.-S. Lam, and R. P. Martins, “A 0.5-V Supply, 36 nW Bandgap Reference With 42 $ppm^{o}C$ Average Temperature Coefficient Within -40°C to 120°C,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 11, pp. 3656–3669, 2020.

[42] K. Nagaraj, T. Viswanathan, K. Singhal, and J. Vlach, “Switched-capacitor circuits with reduced sensitivity to amplifier gain,” *IEEE Transactions on Circuits and Systems*, vol. 34, no. 5, pp. 571–574, 1987.

[43] L. Enthoven, “Cryogenic DAC for the Biasing of Spin Qubits,” Master’s thesis, TU Delft, the Netherlands, 2020.

[44] W. M. C. Sansen, *Analog Design Essentials*. Springer New York, NY, 2006.

[45] L. Yao, M. Steyaert, and W. Sansen, “A 0.8-V, 8-/spl mu/W, CMOS OTA with 50-dB gain and 1.2-MHz GBW in 18-pF load,” in *ESSCIRC 2004 - 29th European Solid-State Circuits Conference (IEEE Cat. No.03EX705)*, 2003, pp. 297–300.

[46] T.-H. Lin, C.-K. Wu, and M.-C. Tsai, “A 0.8-V 0.25-mW Current-Mirror OTA With 160-MHz GBW in 0.18-μm CMOS,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 2, pp. 131–135, 2007.

[47] O. Charlon and W. Redman-White, “Ultra high-compliance CMOS current mirrors for low voltage charge pumps and references,” in *Proceedings of the 30th European Solid-State Circuits Conference*, 2004, pp. 227–230.

[48] M. Mehrpoo, B. Patra, J. Gong, J. van Dijk, H. Homulle, G. Kiene, A. Vladimirescu, F. Sebastiano, E. Charbon, and M. Babaie, “Benefits and Challenges of Designing Cryogenic CMOS RF Circuits for Quantum Computers,” in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–5.