Design of a $-40°C \sim 150°C$ 5σ inaccuracy of $\pm 0.5\%$ Bandgap Reference voltage source

Zitao Zhu

Delft University of Technology

Design of a $-40°C \sim 150°C$ 5σ inaccuracy of $\pm 0.5\%$ Bandgap Reference voltage source

Master's Thesis

Zitao Zhu (5399734)

August 14, 2023

Contents

Acknowledgments

This project has been made possible and successful because of the help and support of many people. I would like to express my gratitude to all of you.

Firstly, I would like to thank my daily mentor, Niraj Kummar, for his frequently discussing and guidance. I would not be able to fulfill the task in time with a good result without his help. I also want to thank my colleagues, Xin Ding and Hui-yu Chen, for helping me get familiar with the tools. And Dave Kroekenstoel for keeping asking constructive questions.

Secondly, I would like to thank my supervisors from TUD, Anton Montagne and Chris Verhoeven. Besides the knowledge, the most precious gift they gave me was thinking in a structured way. This helps me with getting well with all the things in my whole life.

Thirdly, I would like to thank my friends. They all helped me during my hard time and tried their best to help me mentally. Heqi, Jiang, Junwei, Xinyu, Tianyu, Changxu, Huanqiang, Jialing, Yuhao, Xiaomeng, Yige, Yuele, Shenyang, Jie, Fengwei, Qi, Jida, Chenyao, Haoyuan, Chenyan, Jiarui, Yue, Hanshu, Shixun, Weikang, Hongsen, Fanzhe, Pras, and Kyriakos.

Last but most importantly, I want to thank my family for supporting me. I am very happy that I did make my family being proud. I want to thank them for trusting me for so many years even though I did not study well when I was a kid. I would not be able to become me without their love throughout my life.

I also want to record my two understandings of being an engineer. For myself, it is all about having an eternal life: by contributing my efforts to the techniques and letting my design be part of history, that is the way to live forever; As my supervisors used to tell me: We left something here when we left. For the engineering itself: The sage, residing between heaven and earth, stands as the precursor of all beings, observing the opening and closing of yin and yang to give life to all things and understanding the gateway of existence and demise.

To commemorate my beloved grandmother, the first engineer who taught me about the attitude towards life and work.

Z for Impedance and τ for time constant.
$Z\tau$

Abstract

This report presents a bandgap reference voltage source that achieves 5-sigma Inaccuracy of $\pm 0.5\%$ from -40C to 150C under 16FFC process. This is the first time 16*nm* techniques are used in automotive products and the first time trying to realize analog circuits in such a process for in-vehicle network purposes. The report points to good behavior with only a small area and considerable power. It also proves that applying chopping to the circuit does not increase the area.

1 Introduction

1.1 Background

NXP Semiconductor is a company that mainly focuses on in-vehicle network chips. They Keep trying to push the products to the next level by keeping realizing the chips on the latest semiconductor process. In analog integrated circuit design, many basic blocks must be realized at the beginning of a series of analog integrated circuit products. One of the important blocks is a voltage source that provides a voltage level that works as the reference voltage source. Due to the process and mismatch variation during manufacturing, designing such a circuit is hard, especially in the in-vehicle network field. Because in-vehicle network products require a high yield rate for safety requirements, normally in-vehicle network integrated circuits would require 5σ for yield rate, around 99.99997%. Such variation problems are normally expected to be removed by trimming the circuit at a reference temperature. Besides manufacturing, the environment's temperature is also a big problem for the circuit. In-vehicle network chips are expected to function under $-40°C \sim 150°C$. The idea of using the bandgap of silicon is a popular way of realizing such a circuit.

1.2 Current occasion

In the current in-vehicle network chips in NXP semiconductors, an existing design only achieves $\pm 1.25\%$ within the same yield and temperature range. And the circuit was realized based on a $40nm$ technique called C040GF. Now NXP Semiconductor wants to produce better-performance products by moving to a $16nm$ technique called 16FFC. And in the next generation of in-vehicle network chips, the top level requires a higher accuracy for the reference voltage source. So it is preferred to design a voltage reference source that meets the accuracy requirement within the yield and temperature range. And it is preferred to realize such a circuit in 16FFC to prepare for the next generation of chips.

1.3 Research Questions

To summarize, this thesis focuses on the following problem:

Is it possible to design a bandgap reference voltage source that achieves 5-sigma Inaccuracy of $\pm 0.5\%$ from -40C to 150C under the C040GF or 16FFC process? If possible, how much is the cost of the area and power?

1.4 Requirement list

Besides the requirements mentioned in the report's title, some other requirements could be mentioned initially. Note that most of them are soft requirements, which means that as long as the result is roughly at the level, the result will be accepted. The way of testing is also listed. As shown in table 1.

1.5 Report structure

The project is split into four levels and two phases for each level, as shown in Figure 1.1

Table 1: Requirement list for the project

requirement	value	hard or soft?	way of testing
5σ inaccuracy	$< \pm 0.5\%$	hard	checking 5σ worst case under $-40^{\circ}C \sim 150^{\circ}C$
temperature range	$-40^{\circ}C \sim 150^{\circ}C$	hard	
supply voltage range	$1.62V \sim 1.98V$	hard	apply the post-trim result to different supply voltage
output integrated noise($10kHz \sim 10MHz$)	$< 500\mu V_{rms}$	soft	noise simulation
PSR($20log_{10}(\frac{V_{ref}}{V_{out,AC}})$ in $1Hz \sim 1GHz$)	$> 25dB$	soft	AC simulation
power	$< 0.18mW$	soft	DC simulation
area	$< 0.3mm^2$	soft	general floor plan
the temperature at which trimming is performed	$125^{\circ}C$	hard	trim the circuit only at this temperature
settling time	$< 200\mu s$	soft	time doamin simulation

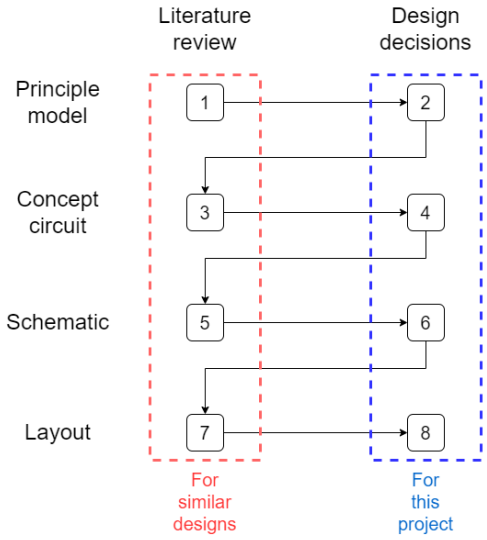

Figure 1.1: Project structure

The four levels: principle model, concept circuit, schematic, and layout. The principle model level uses math models to generate the wanted output and discuss the nonidealities of real devices in a math way. With the nonidealities being analyzed, the project moves to the next level: using a concept circuit to connect the devices and generate the wanted output with a circuit. Note that at this level the controllers are still ideal. The third step is the schematic level, which focuses on realizing the controllers with transistors. The other ideal blocks will be realized at this level as well. A general layout exploration is needed after ensuring the schematic meets the requirement list. Due to the complexity of 16FFC, a commercial-level layout is not required(this would take several extra months to fulfill). The points that need to be kept an eye on will be discussed.

There are two phases for each level: the literature review(Section 2) and the design decisions(3). Design is a ”fill in blanks” job. The literature review is the phase that sets up a structure for such blanks by doing theoretical analysis and listing the boundaries. And base on the theoretical analysis, the design decisions will be made. The phase will also show the result at each level related to the design decisions. Thus the information in the literature review will be useful for all similar designs no matter which process is used. It can be a guild for all similar projects in the future. The design decision section will fully focus on the process and result especially in this project.

2 Literature Review

In this chapter, the required knowledge and the theoretical framework will be discussed. This chapter follows the design flow of a bandgap reference voltage source, focusing on the error sources from design and how they contribute to the output—starting from the principle model level, the concept circuit level, the schematic level, and the layout level. The detailed design decisions will be made and shown in Section 3. The circuit behavior of each design step will be there as well.

2.1 Principle model level

The principle model level starts with the operating principle of the bandgap reference voltage. At this level, the Bipolar junction transistors(BJT) are the focused points while analyzing. The reason is that they are the components that form the circuit's operation in principle. This section discusses how the BJT creates the reference voltage at the output and the error from the BJT. After discussing each error source, a conclusion that supports designing will be delivered. A design example will be raised to verify the analysis, and the estimated result will be compared with the Monte Carlo result from the simulation tool.

2.1.1 Bipolar junction transistor model

When a PN-junction is forward biased, the relation between the current flows through and the voltage applied to both sides can be expressed as follow:

$$I = I_s \cdot (e^{\frac{V}{V_t}} - 1) \tag{2.1}$$

If the current injected into the PN-junction is constant among different temperatures. Then the voltage generated by the PN-junction can be expressed as follow:

$$V = \frac{k \cdot T}{q} \cdot ln(\frac{I}{I_s}) \tag{2.2}$$

Unfortunately, due to the nonlinearity of the device, in practice, it's actually $V \propto T^{1.1}$. A better way to realize $V \propto T$ is to use a diode-connected Bipolar junction transistor(BJT) to replace such PN-junction. When it comes to a diode-connected BJT, the relation between the current flows in the Collector and the V_{BE} across its Base and Emitter can be expressed as follows:

$$I_c = I_s \cdot (e^{\frac{V_{BE}}{V_t}} - 1) \tag{2.3}$$

Since even for tiny Base-Emitter voltages, the exponential term is already quite significant. Thus the -1 term is negligible. By rewriting the equation, the equitation can be rewritten in (2.4):

$$V_{BE} = \frac{k \cdot T}{q} \cdot ln(\frac{I_c}{I_s}) \tag{2.4}$$

In which I_c is the collector bias current, I_s is the saturation current, q is the electron charge($1.6 \cdot 10^{-19}C$), T is the absolute temperature(T) and k is Boltzmann constant($1.38 \cdot 10^{23}J/K$). In this equation, the k and q are constants who does not vary with the temperature. I_c is a design parameter that can be a constant or proportional to the temperature, as expressed in (2.5). The temperature dependency of I_s is expressed in(2.6)[1].

$$I_c(T) = I_c(T_r) \cdot (\frac{T}{T_r})^{\theta} \tag{2.5}$$

$\theta = 0$ for constant I_c and $\theta = 1$ for 1^{st} order temperature proportional(PTAT) I_c.

$$I_s(T) = C \cdot (\frac{T}{T_r})^{\eta} \cdot e^{-\frac{E_{GO}}{kT}} \tag{2.6}$$

The η is the order of temperature dependency defined by the process. And $E_{GO} \approx 1.206eV$ is the extrapolated bandgap energy of silicon at $T = 0K$. And C is a constant also defined by process.

With combining the (2.4), (2.5), and (2.6), a new expression of V_{BE} can be delivered:

$$V_{BE}(T) = \frac{E_{GO}}{q} + \frac{kT}{q} ln(\frac{I_c(T_r)}{C}) - (\eta - \theta)\frac{kT}{q} ln(\frac{T}{T_r}) \tag{2.7}$$

To make this a simpler equation for design, apply Taylor expansion on (2.7) to get a first-order model for V_{BE}.

$$V_{BE}(T) = V_{BE}(T_r) + [V_{BE}(T_r) - \frac{E_{GO}}{q} - \frac{kT_r}{q}(\eta - \theta)] \cdot \frac{T - T_r}{T_r} \tag{2.8}$$

Now with introducing a new term to represent the temperature non-dependent part:

$$V_G = \frac{E_{GO}}{q} + \frac{kT_r}{q}(\eta - \theta) \tag{2.9}$$

Then (2.8) turns into a first-order line in (2.10) defined by two points: $1.V_{BE} = V_G$ at $T = 0K$; $2.V_{BE} = V_{BE}(T_r)$ at $T = T_r$. As shown in Figure 2.1. By define the V-T relationship with only two points helps understanding and designing much easier.

$$\begin{aligned} V_{BE}(T) &= V_{BE}(T_r) + [V_{BE}(T_r - V_G)] \cdot \frac{T - T_r}{T_r} \\ &= V_G - (V_G - V_{BE}(T_r)) \cdot \frac{T}{T_r} \end{aligned} \tag{2.10}$$

Since V_G is a constant defined by the temperature dependency process and order of I_c, and usually I_c is designed as a first-order PTAT current. Therefore, V_G can be considered a constant for the design in the same process. Thus, by only creating the V_{BE} at the reference temperature $T = T_r$, The whole V_{BE} curve can be defined. And the rest part can be taken as a curvature error(discussed in the following section). This also answers why using diode-connected BJT instead of a diode. This is because the diodes' V-T curve is not as first-order as the diode-connected BJT. Usually, the V-T curve of a BJT has a slightly higher order.

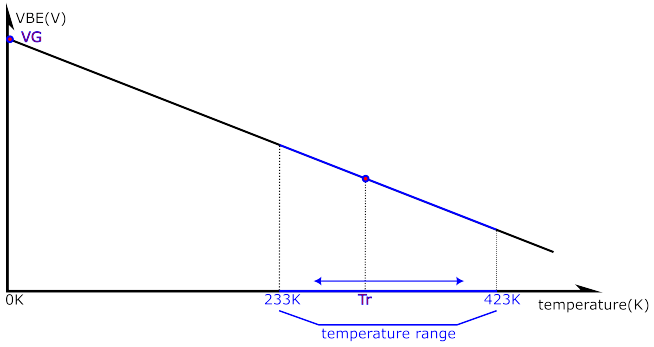

Figure 2.1: V_{BE} can be defined by two points

2.1.2 Forming Temperature non-dependent output

With the complimentary to absolute temperature(CTAT) V_{BE}, two methods exist to realize a constant output.

CTAT-CTAT Using two CTAT sources with different $V_{BE}(T_r)$ to generate two various V_{BE} curves with different slopes. By amplifying one of the curves and making both curves' slopes equal. A constant output can be delivered by taking the difference between those two V_{BE} curves. As shown in Figure 2.2.

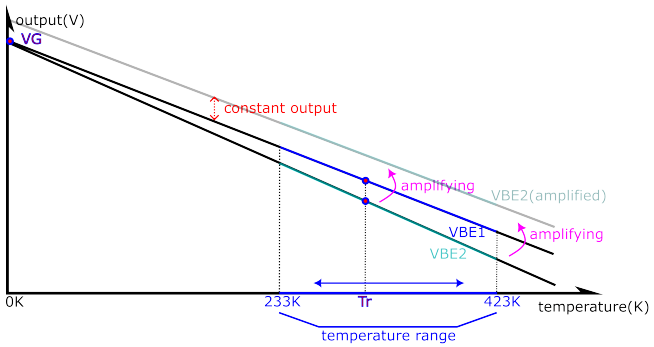

Figure 2.2: CTAT-CTAT

The equation for expressing this principle model can be written as (2.11). Please note that a_1, a_2 can have any sign; this allows all the possibility of the principle model.

$$V_{output} = a_1 V_{BE1} + a_2 V_{BE2} \tag{2.11}$$

PTAT+CTAT Another method uses a proportional to absolute temperature(PTAT) voltage to mitigate the dropping of V_{BE}. By having the same absolute value in slope through amplifying, the sum of two voltages can deliver a constant output, as shown in Figure 2.3. Usually, this voltage is created by the difference of two V_{BE}, as described in (2.12)

$$\Delta V_{BE} = V_{BE1} - V_{BE2} = \frac{kT}{q} \cdot ln(\frac{Ic_1}{Is_1} \cdot \frac{Is_2}{Ic_2}) \tag{2.12}$$

Due to the $\frac{kT}{q}$ in the ΔV_{BE}, this PTAT source starts from $\Delta V_{BE} = 0$ at $T = 0K$. Thus, similar to the definition of V_{BE}. The PTAT source is can be defined by two points: $1.\Delta V_{BE} = 0$ at $T = 0K$; $2.\Delta V_{BE} = \Delta V_{BE}(T_r)$ at $T = 0K$.

After getting the ΔV_{BE}, the output of the principle model becomes (2.13). Please note that a_1 and a_2 can have any sign; this allows all the possibilities of the principle model.

$$V_{output} = a_1(V_{BE1} - V_{BE2}) + a_2 V_{BE1} \tag{2.13}$$

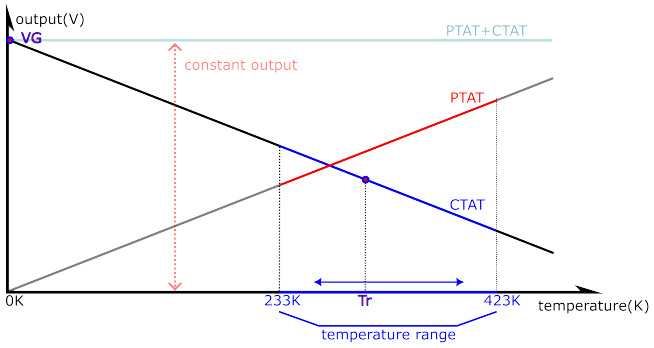

Figure 2.3: PTAT+CTAT

Comparison of two methods The exciting thing is that, by allowing the possibility for having positive or negative signs in a_1, a_2. In some conditions, the PTAT+CTAT can become CTAT-CTAT. (e.g. $a_1 = -1.5, a_2 = 1, V_{BE2} > V_{BE1}$). This means that by not defining the signs of a_1, a_2, analyzing all cases on PTAT+CTAT can cover the case of CTAT-CTAT.

Besides the convenience of PTAT+CTAT, the other reason for focusing on that is the behavior varying while manufacturing. More components always lead to more variation during manufacturing. Thus more straightforward structure always promises better behavior after manufacturing. PTAT+CTAT requires only two BJT to realize the constants output, and CTAT-CTAT requires an additional PTAT current source, which requires another two BJT to form the ΔV_{BE} and makes the structure more complex. For the same reason, the idea of realizing PTAT+CTAT in (2.14) is not in the considering range.

$$V_{output} = a_1(V_{BE1} - V_{BE2}) + a_2 V_{BE3} \tag{2.14}$$

Insert the V_{BE} model into the method By inserting (2.10) into (2.13). It comes to (2.15).

$$\begin{aligned} V_{output} &= a_1(V_{BE1} - V_{BE2}) + a_2 V_{BE1} \\ &= a_2 V_G + \frac{(a_1+a_2)V_{BE1}(T_r) - a_1 V_{BE2}(T_r) - a_2 V_G}{T_r} \cdot T \end{aligned} \tag{2.15}$$

For a demand of constant V_{output}, it is wished to have a constant non-zero part and a zero-coefficient temperature-dependent part. the coefficients a_1, a_2 need to be designed to fulfill the condition in (2.17).

$$\begin{cases} a_2 V_G = V_{ref} \\ \frac{(a_1+a_2)V_{BE1}(T_r) - a_1 V_{BE2}(T_r) - a_2 V_G}{T_r} = 0 \end{cases} \tag{2.16}$$

By rewriting the (2.16), the definition of a_1 and a_2 is delivered in (2.17).

$$\begin{cases} a_2 = \frac{V_{ref}}{V_G} \\ a_1 = \frac{V_{ref} - \frac{V_{ref}}{V_G} V_{BE1}(T_r)}{V_{BE1}(T_r) - V_{BE2}(T_r)} \end{cases} \tag{2.17}$$

From (2.17), the duty of a_1 and a_2 can be told. a_2 is the coefficient that defines the constant output, which will be taken as the reference voltage V_{ref}. a_1 is the coefficient that balances the PTAT and CTAT source.

As discussed in the previous section, the $V_{BE}(T)$ curve can be defined by $V_{BE}(T_r)$, and $V_BE(T_r)$ is defined by $I_s(T_r)$ and $I_c(T_r)$ through (2.4) at $T = T_r$. With two BJT, there are four design parameters. Minimizing the number of design parameters for a trade-off analysis is always good. So two new design parameters defined at $T = T_r$ are introduced in (2.18).

$$\begin{cases} I_{budget} = I_{c1} + I_{c2} \\ I_{c1} = I_{c_{ratio}} \cdot I_{budget} \end{cases} \tag{2.18}$$

As mentioned in the beginning, the BJTs are biased with a first-order PTAT current. Usually, this PTAT current is generated by ΔV_{BE} from (2.12). Thus the sum of biasing current I_{budget} can also be defined in the same way, as shown in (2.19)

$$I_{budget}(T) = I_{budget}(T_r)\frac{T}{T_r} \tag{2.19}$$

This means that, with the I_{budget} defined at $T = T_r$, the $I_{c_{ratio}}$ defined for splitting the current into I_{c1} and I_{c2}, and knowing the I_{s1} and I_{s2} at $T = T_r$. The bias condition(V_{BE}) of two BJT at $T = T_r$ can be defined. Thus a_1 is defined by V_{ref}, $I_{budget}(T_r)$, $I_{c_{ratio}}$, $I_{s1}(T_r)$, and $I_{s2}(T_r)$. And a_2 is defined by V_{ref}.

2.1.3 Error sources in BJT

Now, with a good definition of all the components in the forming of V_{output} in (2.13). Then, it's time to analyze the impact of error sources in BJT. There are five primary error sources in BJT[2][1]: saturation current I_s, current gain β, Base resistance r_b, and curvature of V_{BE}. The error sources will be discussed separately in the following sections Focusing on how they participate in the transfer function to the V_{output} and the method to mitigate them. In the end, there will be a section expressing an example of an analysis of PNP BJTs in the C040GF process.

Before starting with the analysis of each error source, a definition of error is introduced. From (2.15), it can be told that all the BJT-based errors participate in the forming of V_{output} via the temperature-dependent part. Thus, the errors can be split from the (2.15), and it comes to (2.20).

$$\begin{aligned} V_{error} &= \frac{(a_1+a_2)V_{BE1}(T_r) - a_1V_{BE2}(T_r) - a_2V_G}{T_r} \cdot T \\ &= \frac{a_1(V_{BE1}(T_r) - V_{BE2}(T_r)) + a_2V_{BE1}(T_r) - a_2V_G}{T_r} \cdot T \end{aligned} \tag{2.20}$$

This error is a straight curve starting from $0V$ at $T = 0K$. Ideally, with a given designed a_2, the slope of this curve is kept as zero by a_1. However, when error sources from BJT influence the $V_{BE}(T_r)$, the balance between V_{BE1} and V_{BE2} is broken, then a_1 needs to be adjusted to re-balance them. In practice, this adjustment is realized by trimming.

According to (2.17), both a_1 and a_2 has V_{ref} in them, By dividing a_1 and a_2 with V_{ref}, $\frac{a_1}{V_{ref}}$ and $\frac{a_2}{V_{ref}}$ can be V_{ref}-independent, as shown in 2.21.

$$\begin{cases} \frac{a_2}{V_{ref}} = \frac{1}{V_G} \\ \frac{a_1}{V_{ref}} = \frac{1-\frac{1}{V_G}V_{BE1}(T_r)}{V_{BE1}(T_r)-V_{BE2}(T_r)} \end{cases} \tag{2.21}$$

Then apply this to (2.20), it comes to (2.22).

$$\begin{aligned}\frac{V_{error}}{V_{ref}} &= \frac{\frac{a_1}{V_ref}[V_{BE1}(T_r) - V_{BE2}(T_r)] + \frac{V_{BE1}(T_r)}{V_G} - 1}{T_r} \cdot T \\ &= \frac{\frac{a_1}{V_ref}[V_{T_r} \cdot ln(\frac{I_{c1}}{I_{s1}} \cdot \frac{I_{s2}}{I_{c2}})] + \frac{1}{V_G}[V_{T_r} \cdot ln(\frac{Ic_1}{I_{s1}})] - 1}{T_r} \cdot T\end{aligned} \tag{2.22}$$

Now all the components in $\frac{V_{error}}{V_{ref}}$ are V_{ref}-independent, this means that with the analysis on $\frac{V_{error}}{V_{ref}}$, the amount of errors refer to V_{output} can be easily worked out by $\frac{V_{error}}{V_{ref}} \cdot V_{ref}$. Which would save much time and effort.

2.1.4 Error source: saturation current

Introduction of error source: saturation current Two errors lead to variation in saturation current, variation during processing, and mismatch between I_{s1} and I_{s2} during manufacturing. These two errors are represented by the two normalized parameters, $error_{process}$ and $error_{mismatch}$ in (2.23).

$$\begin{cases} I_{s1,errored} = I_{s1,nominal} \cdot (1 + error_{process}) \\ I_{s2,errored} = I_{s2,nominal} \cdot (1 + error_{process}) \cdot (1 + error_{mismatch}) \end{cases} \tag{2.23}$$

The process variation of saturation current is caused by the spread of Base doping and geometric area. Thus, these normalized parameters are not temperature-dependent.

while modeling the $V_{BE}(T)$, the $V_{BE}(T_r)$ at $T = T_r$ is needed. And with knowing the bias current I_c and I_s, the $V_{BE}(T_r)$ is defined. An error in the saturation current makes the $V_{BE}(T)$ curve change. Then (2.22) becomes (2.24).

$$\frac{V_{error}}{V_{ref}} = \frac{\frac{a_1}{V_ref}[V_{T_r} \cdot ln(\frac{I_{c1}}{I_{s1}} \cdot \frac{I_{s2}}{I_{c2}} \cdot (1 + error_{mismatch}))] + \frac{1}{V_G}[V_{T_r} \cdot ln(\frac{Ic_1}{I_{s1}} \cdot \frac{1}{1+error_{process}})] - 1}{T_r} \cdot T \tag{2.24}$$

Now plot this error in Figure 2.4.

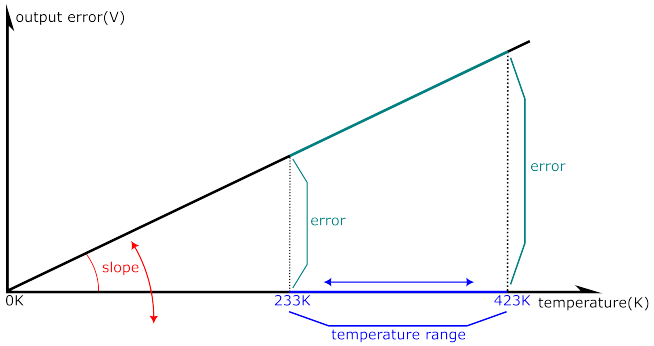

Figure 2.4: Saturation current caused error at V_{output}

An example of the impact of the worst-case process and mismatch variation is also shown in Figure 2.5. In the figure, n and p represent the least or most value for $I_{s_{e1}}(error_{process})$ and $I_{s_{e2}}(error_{mismatch})$.

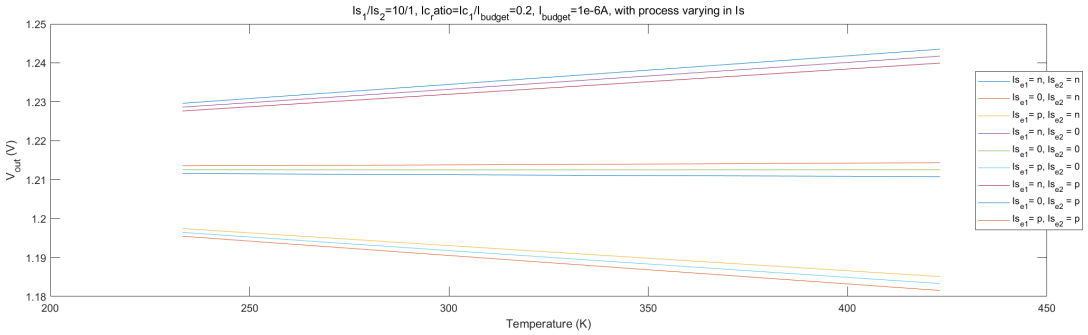

Figure 2.5: An example of the impact of I_s's process and mismatch variation

Mitigation of error source: saturation current The slope part is the coefficient in (2.24), which will be a constant while the temperature changes. This results in a first-order error. This slope can be zero again by adjusting the a_1 to rebuild the balance between V_{BE1} and V_{BE2} since this error starts from $(0K, 0V)$. An adjusting of a_1 at only $T = T_r$ is enough for the correcting.

It is always preferred to adjust the a_1 as less(by %) as possible. This can be realized by attenuating the influence of the errors at the output. That is decreasing the gain of their transfer function to the output. From (2.24), it can be told that: With the same amount of process variation and mismatch.1.With a lower $|\frac{a_1}{V_{ref}}|$ at nominal case, the impact of mismatch can be lower; 2.There is nothing that can be done to minimize the impact of process variation since its coefficient $\frac{V_{Tr}}{V_G}$ is defined by nature and process.

For a lower $|\frac{a_1}{V_{ref}}|$, a extended version of (2.21) is rewritten in (2.25)

$$\frac{a_1}{V_{ref}} = \frac{1 - \frac{1}{V_G} V_{T_r} \cdot ln(\frac{I_{c_{ratio}} \cdot I_{budget}}{I_{s1}})}{V_{T_r} \cdot ln(\frac{I_{c_{ratio}} \cdot I_{budget}}{I_{s1}} \cdot \frac{I_{s2}}{(1 - I_{c_{ratio}}) \cdot I_{budget}})} \tag{2.25}$$

By increasing $\frac{I_{s1}}{I_{s2}}$ and decreasing $I_{c_{ratio}}$, the $|\frac{a_1}{V_{ref}]}|$ can be reduced. Figure 2.6 shows an example of a_1 vary with $\frac{I_{s1}}{I_{s2}}$ and $I_{c_{ratio}}$. In this example, $I_{s2} = 3.4e-19A, n = \frac{I_{s2}}{I_{s1}}, V_{ref} = 0.2V$

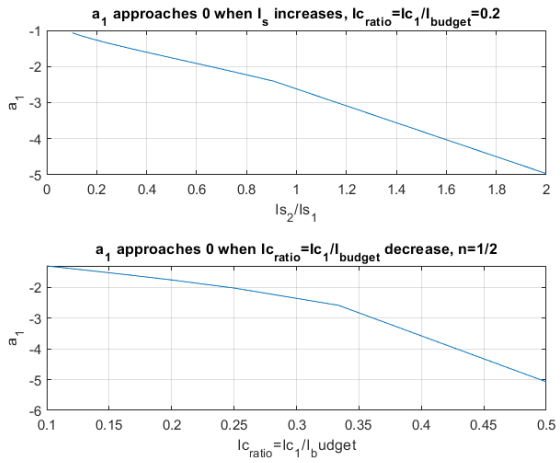

Figure 2.6: An example of a_1 value while varying $\frac{I_{s1}}{I_{s2}}$ or $I_{c_{ratio}} = \frac{I_{c1}}{I_{budget}}$, note that I_{budget} is defined at T_r

This is because: if the V_{ref} is defined, then a_2 is settled. This means that the slope of the CTAT source is settled(a_2 is the coefficient of the CTAT source). For balancing the sum, the required slope of the PTAT source is also settled. Since the PTAT curve is defined by $(0V, 0K)$ and $(\Delta V_{BE}, T_r)$, bigger I_{s1} and less $I_{c_ratio} = \frac{I_{c1}}{I_{budget}}$ leads to a bigger $\Delta V_{BE}(T_r)$ and bigger slope for PTAT curve, this would requires less $|a_1|$ to produce the required slope. Please keep this conclusion in mind; it will also be used in the following sections.

2.1.5 Error source: current gain

Introduction of error source: current gain The current gain of a BJT is defined as the ratio between the Collector current and Base current, as shown in (2.26):

$$\beta = \frac{I_c}{I_b} \tag{2.26}$$

Taking the diode-connected PNP BJT as an example(Figure 2.7), the Base and Collector are connected to achieve a condition of $V_{BC} = 0$; a current source is biasing the BJT by controlling its Emitter current.

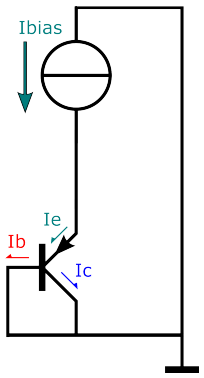

Figure 2.7: Collector current and Base current split the Emitter current

With the relationship expressed in (2.26), now the Base current and Collector can be delivered in :

$$\begin{cases} I_b = \frac{1}{1+\beta} \cdot I_{bias} \\ I_c = \frac{\beta}{1+\beta} \cdot I_{bias} \end{cases} \tag{2.27}$$

In the past, β was a pretty high value. This makes I_b negligible, which allows designers assuming $I_c \approx I_{bias}$. But in the modern CMOS process, β is a value around 1.5 at room temperature, which makes I_b no longer negligible. This results in a new expression for V_{BE}, as shown in (2.28)

$$V_{BE} = \frac{kT}{q} ln(\frac{I_{bias} \cdot \frac{\beta}{1+\beta}}{I_s}) \tag{2.28}$$

Even worse, the current gain is a temperature-dependent value, and the dependency is different in different processes. This section uses a first-order PTAT β as an example, as shown in Figure 2.8. β is also biasing-current-density-dependent; in some processes, this leads to a different current gain between two BJTs. The width of the Emitter also leads to different current gains. Besides the non-ideality brought by the current gain in the nominal case, it also has process and mismatch variation.

The example of analysis in this section is based on C040GF, which has a low biasing-current-density dependency. The rest above errors will be discussed in the following paragraphs.

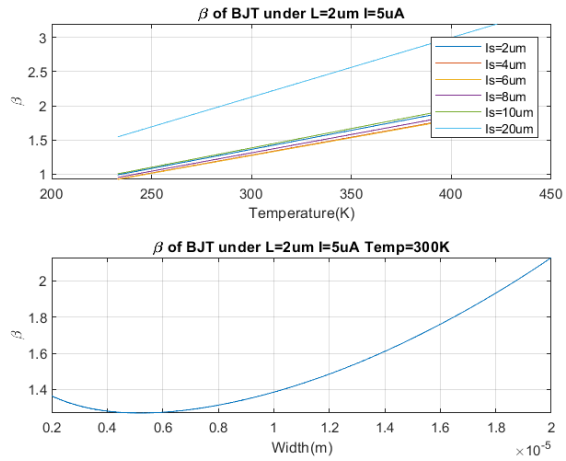

Figure 2.8: β vary with temperature and Emitter width(with least Emitter length($2\mu m$))

As shown in Figure 2.9, with a $\frac{\beta(T)}{1+\beta(T)}$ applied to the V_{BE}, V_{BE} is lower than expectation. For different temperatures, there will be a different β. Thus the $V_{BE}(T_r)$ will be established under different $\frac{\beta}{1+\beta}$. For higher temperatures, β is bigger, $\frac{\beta}{1+\beta}$ is closer to 1. Thus the V_{BE} models for the higher temperatures are closer to the ideal one. By combining the $V_{BE}(T)$ at different temperatures, the real $V_{BE}(T)$ becomes a slightly curved line modeled by multiple straight line models at different temperatures. The result of this model is close to the results in cadence under low biasing current density occasion, which is good enough for designing.

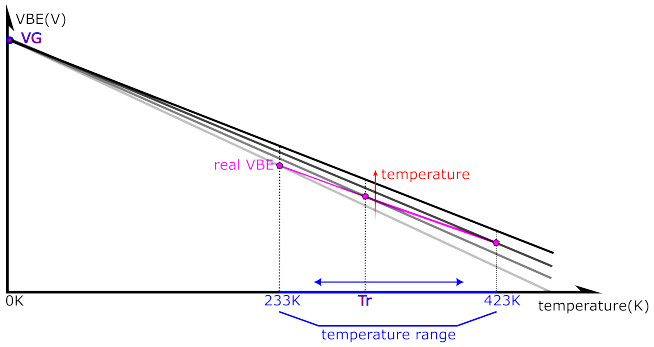

Figure 2.9: β changes the V_{BE} model for each different temperature

Influence of temperature-dependent current gain First, consider the current gain in the nominal case without any process and mismatch variation. Out of the consideration of area and mismatch, choose the BJTs to have $W_{Emitter} = 2\mu m, L_{Emitter} = 2\mu m$. This leads to a new expression for the error, as shown in (2.29):

$$\frac{V_{error}}{V_{ref}} = \frac{\frac{a_1}{V_ref}[V_{T_r} \cdot ln(\frac{I_{c1}}{I_{s1}} \cdot \frac{I_{s2}}{I_{c2}})] + \frac{1}{V_G}[V_{T_r} \cdot ln(\frac{I_{c1}}{I_{s1}} \cdot \frac{\beta}{1+\beta})] - 1}{T_r} \cdot T \tag{2.29}$$

Different from the saturation mismatch, this error has a temperature-dependent slope, as shown in Figure 2.10. Note that $ln(\frac{\beta}{1+\beta})$ shall lead to a negative error shift in slope. For easier understanding, the curve is drawn in the first quadrant, and the sketched curve should be the absolute value of the error.

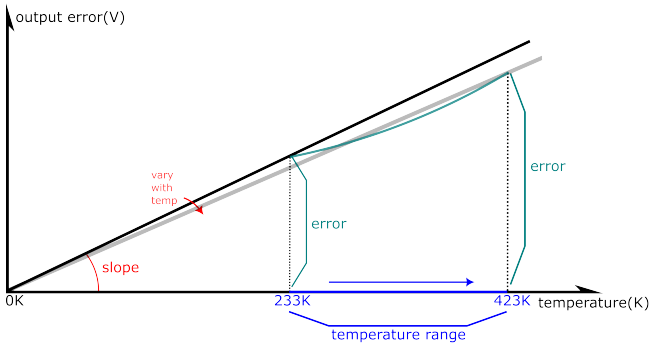

Figure 2.10: β leads to an error which has a temperature-dependent slope

Since the way of modeling β is different for each process, this paragraph will skip that detail. It is for sure that as the temperature keeps rising, β keeps rising as well. Thus $\frac{\beta}{1+\beta}$ approaches 1 when the temperature rises. This means that the slope for the error keeps dropping with the temperature rising. Thus current gain of BJT leads to a high-order error.

Another way to understand this phenomenon is: The current gain makes the Collector current lower than expected. This effect does not occur in the $\frac{I_{c1}}{I_{c2}}$ but only influences the CTAT source($V_{BE1}(T)$) who forms output. According to Figure 2.9, the real V_{BE} has a smaller slope. This means that it will no longer fully cancel the PTAT part's slope in the output; some parts of the PTAT slope will be kept in the output. Since the real V_{BE} curve is a high-order line now, this is also the reason why the error is a PTAT high-order curve.

Mitigation of temperature-dependent current gain According to (2.29), a_1 is part of the slop of the error. As a tool for balancing the output, a_1 should only be allowed to be adjusted once for all temperatures. Thus the slope of error can only be changed once for all temperatures. Changing the error's slope can cancel all the error's first-order parts. The best case a_1 can adjust to the best case by letting $V_{output}(233K) = V_{output}(423K)$, as shown in Figure 2.11. This also means at least two-temperature trimming is required because it is needed to check the tendency of the error now.

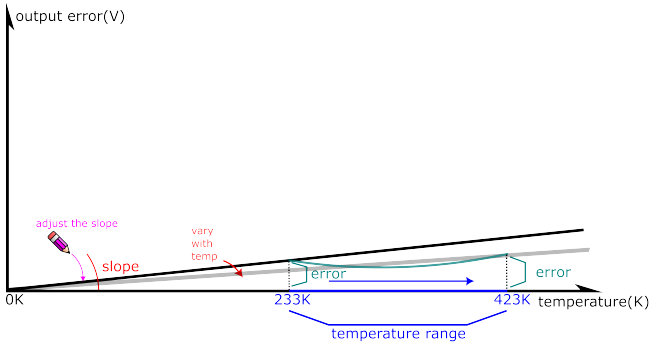

Figure 2.11: Adjusting a_1 can remove all the first-order parts of the error

Note that the rest high-order part of the error does not vary around $error = 0$. This will not be a problem since now the reference voltage can be taken as $V_{ref,ideal} + average[min(error), max(error)]$, and the result of this error is negligible($0.05mV$ variation for $V_{output} \approx 1250mV$). The detail will be shown in the example section.

Influence of current gain's process& mismatch variation With the conclusion from last paragraph, now $V_{ref,ideal} + average[min(error), max(error)]$ can be taken as the standard performance at nominal case. Then it's time to add the process and mismatch variation into the analysis as shown in (2.30). The process and mismatch errors are added to the current gain model, similar to the saturation current variation.

$$\begin{cases} \frac{V_{error}}{V_{ref}} = \frac{\frac{a_1}{V_{ref}}[V_{Tr} \cdot ln(\frac{I_{c1} \cdot servo_{\beta_1}}{I_{s1}} \cdot \frac{I_{s2}}{I_{c2} \cdot servo_{\beta_2}} \cdot)] + \frac{1}{V_G}[V_{Tr} \cdot ln(\frac{Ic_1 \cdot servo_{\beta_1}}{I_{s1}} \cdot)] - 1}{T_r} \cdot T \\ servo_{\beta_1} = \frac{\beta_1(T) \cdot (1 + error_{process})}{1 + \beta_1(T) \cdot (1 + error_{process})} \\ servo_{\beta_2} = \frac{\beta_2(T) \cdot (1 + error_{process}) \cdot (1 + error_{mismatch})}{1 + \beta_2(T) \cdot (1 + error_{process})(1 + error_{mismatch})} \end{cases} \tag{2.30}$$

With the a_1 designed considering the effect caused by β at the nominal case, the process and mismatch variation can amplify or attenuate the β. Considering all the worst cases, an example of the range of various results is shown in Figure 2.12. In the figure, n and p represent the least or most value for $\beta_{process_1}(error_{process})$ and $\beta_{process_2}(error_{mismatch})$.

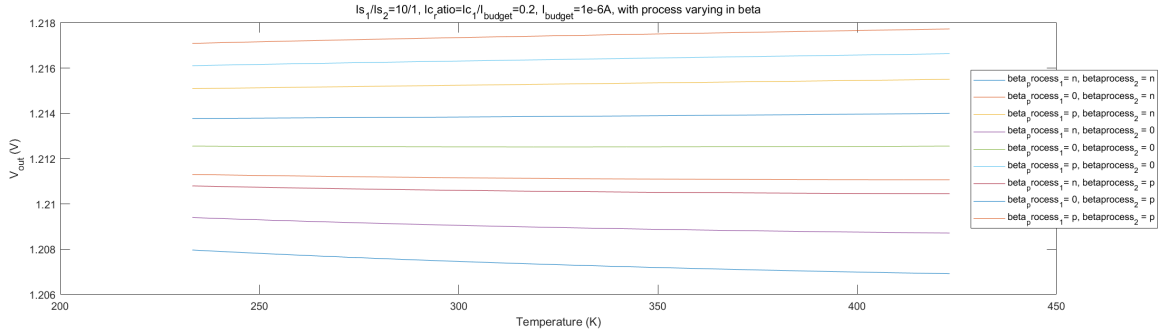

Figure 2.12: An example of the impact of β's process and mismatch variation

According to Figure 2.9, a bigger β makes CTAT slop closer to ideal case, which is a bigger slope. Smaller β works in the opposite way. While manufacturing, the process variation is usually much bigger than the mismatch variation. This variation amplifies or attenuates the CTAT's slope ($\frac{I_{c1}}{I_{s1}}$ variation is process variation dependent) a lot but only slightly in PTAT's slope($\frac{I_{c1}}{I_{c2}}$ variation is process variation dependent). This is why it can be observed in Figure 2.12 that with a balanced designed a_1 at the nominal case, there is still symmetric variation located around the such nominal case.

Mitigation of current gain's process& mismatch variation To check how much error adjusting a_1 can mitigate, an analysis of the two boundary cases is shown in Figure 2.13. The a_1 of each case is redesigned based to meet $V_{output}(233K) = V_{output}(423K)$. The result in this example shows that there is still around $\pm 0.3\%$ error which is not cancellable. This is the ideal case based on calculation, which means that the after-adjusted result will be worse in practice.

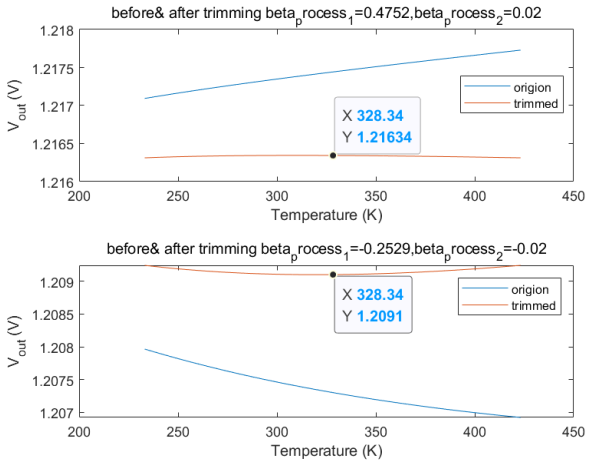

Figure 2.13: An example of the best-adjusted result under worst β's process and mismatch variation

According to (2.30). A conclusion similar to the one in saturation current analysis can be drawn. A smaller $|a_1|$ reduces the gain between the output and the mismatch-variation-leading error. And the coefficient of the process-variation-leading error is not designable if V_{ref} is settled.

Mitigation of error source: current gain As the saying goes: prevention is better than cure. Preventing β from participating in the forming of V_{output} is always better than trying to mitigate that by adjusting a_1.

The first method is making the current way bigger than 1. This will make the variation of $\frac{\beta}{1+\beta} \approx 0$ at all temperatures and all variation cases negligible. The cost of this method is area. In nowadays process, the most efficient way of increasing the current of a BJT is to increase its Emitter width.

The second method is using a follower. By applying A controller to form a follower to supply the Base with the same voltage as the Collector, the BJT is diode connected with its Collector and Base isolated. The required Base current will be supplied by the controller, which is directly from the power supply. Two follower-applying examples are shown in figure 2.14. This idea's key point is taking Collector current as bias current and supplying no matter how much current is required by Base. This method was also used in [1] and [3]. Considering the biasing, it will be easier to implement the follower with an NPN BJT. Then the β caused errors will be replaced by errors caused by followers. If follower-caused errors can be lower, then the proposed idea is valid.

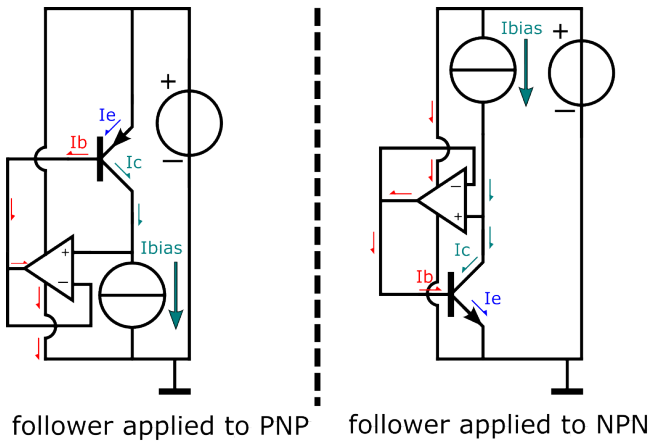

Figure 2.14: Ideal for avoiding current gain's impact

The cost of this method is the limited DC gain of the controller and the offset brought by the controller. Take the NPN case as an example to check the show-stopper value for the DC gain and the offset. As shown in Figure 2.15.

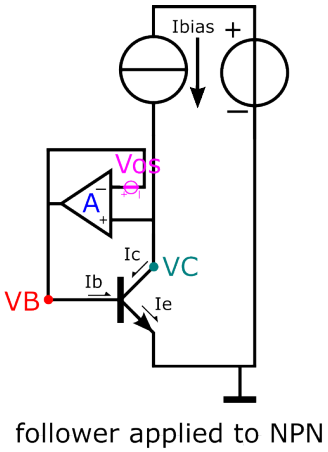

Figure 2.15: Errors introduced by follower

A is the DC gain of the controller used to form the follower, which is wished to be as big as possible. With an accurate I_bias flowing into the Collector, the V_{BE} is defined. The controller requires needs enough V_C to enable such a V_{BE}, as shown in (2.31). This means V_{BC} will have a $\frac{1}{A} \cdot V_{BE}$ error on it. Ideally, V_{BC} is wished to be zero; otherwise, V_{CE} cannot be taken as ideal V_{BE}. At the lowest temperature, $V_{BE} \approx 800mV$, This leads to $\frac{1}{A}800mV$. A controller with more $60dB$ DC gain can make this error $\pm 0.03\%$, which is easy to achieve in modern CMOS processes.

$$V_B = V_C \cdot \frac{A}{1+A} \tag{2.31}$$

V_{os} is the offset between the input pairs, which is also the biggest error source. This will directly result in V_{CE} a mismatch, which makes the V_{CE} not the ideal V_{BE}. According to (2.13), the gain from the error to the output is $a_1 + a_2$ and a_1. And the offset boundaries at the output are the positive and negative versions of their absolute sum, as shown in(2.32).

$$\begin{cases} error_{follower1} = (a_1 + a_2) \cdot V_{os} \\ error_{follower2} = a_1 \cdot V_{os} \\ error_{follower} = \pm|(2 \cdot a_1 + a_2) \cdot V_{os}| \end{cases} \tag{2.32}$$

For $a_2 = 1$, it is easy to achieve $a_1 \approx -5$ without costing too much area. With offset canceling techniques, achieving a $0.5mV$ offset is pretty easy, which leads to a $4.5mV$ error at the output. This is lower than the $8mV$ non-trimmable error caused by the current gain. Besides, this also saves time for trimming. Removing the β from the V_{output} makes two temperature trimming no longer needed. Besides that, due to the different variations of saturation current and current gain, the optimum adjusted a_1 may not be the same. Thus even the technically fully cancellable error may still be left a little. Replacing the formal error with a new one and canceling that new error with another method can make adjustments to cancel more errors.

2.1.6 Error source: Base resistance

Introduction of error source: Base resistance As mentioned in the last section, Base current is not negligible in modern techniques. When the base current keeps increasing, the real V_{BE} will be more than the estimated V_{BE} from (2.4). With increasing the Base current, this difference rises as well. This phenomenon can be represented by a resistor R_B in series with the intrinsic Base, as shown in (2.33) and Figure 2.16. The detail analyzing of controller offset will be in further sections.

$$V_{BE} = \frac{k \cdot T}{q} \cdot ln(\frac{I_c}{I_s}) + I_b \cdot R_B \tag{2.33}$$

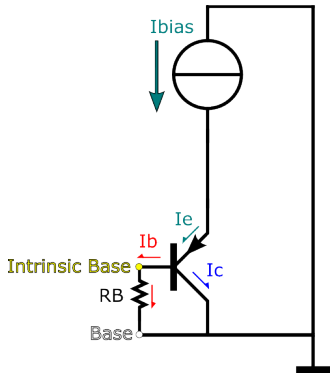

Figure 2.16: Base resistance in series with the intrinsic Base

Considering one of the BJT is a multiple-paralleled version of the other, its base resistance will be multiple times smaller. In the meantime, the bias current in that larger BJT is designed to be low to achieve a big $\frac{I_{c2}}{I_{c1}} \cdot \frac{I_{s1}}{I_{s2}}$(in this formula BJT1 is the big one). Thus only the impact of the Base resistance of the smaller BJT(in the example BJT2) is considered.

According to the (2.13), the term $I_b \cdot R_B$ only occurs in V_{BE2}. Thus the error referring to output can be expressed as (2.34)

$$error = |a_1| \cdot R_B \cdot I_b \tag{2.34}$$

In this equation, a_1 is a constant, R_B is bias-dependent and temperature-dependent, and I_b is temperature dependent. R_B decreases with temperature or bias current increasing ranges between 100Ω and 400Ω(in example process). I_b increases with temperature increasing, but not first order and the slope will be lower than the slope of bias current generated by ΔV_{BE}. This is because $\frac{1}{1+\beta}$ decrease while temperature increasing.

Mitigation of error source: Base resistance There are two methods to mitigate the error brought by Base resistance. The first is to design $|a_1|$ to be as small as possible, lowering the error gain. The other method is to bias the BJT with a low current density. For example, a bias current $I_e \approx 1uA$ is used in some designs. This leads to $V_{R_B} \approx 0.2mV$ and an error of $\pm 0.008\%$, which is negligible. And This current design still promises good performance in other aspects.

2.1.7 Error source: Curvature

Introduction of error source: Curvature Except for the first order part of V_{BE}, there is a high order part, as shown in (2.35).

$$V_{BE}(T) = V_G - [V_G - V_{BE}(T_r)] \cdot \frac{T}{T_r} - XTI \cdot \frac{kT}{q} \cdot ln(\frac{T}{T_r}) + \frac{kT}{q} \cdot ln(\frac{I_c}{I_{c_r}}) \tag{2.35}$$

The first two terms are the first-order model of the V_{BE}. XTI is a model parameter related to carrier mobility [4], I_{c_r} is the Collector current at $T = T_r$. Since $\frac{T}{T_r}$ and $\frac{I_c}{I_{c_r}}$ shall be same for both BJTs, this error will be canceled in $\Delta V_{BE} = V_{BE1} - V_{BE2}$. According to (2.13), the transfer from the error to the output is (2.36)

$$error = |a_2| \cdot [-XTI \cdot \frac{kT}{q} \cdot ln(\frac{T}{T_r}) + \frac{kT}{q} \cdot ln(\frac{I_c}{I_{c_r}})] \tag{2.36}$$

This process-dependent error has nothing to do with the BJT geometric parameters. And in the example process C040GF, this curvature error leads to a variation of around $2.5mV$, which is around $\pm 0.1\%$ when $a_2 = 1, V_{ref} \approx 1.2V$. After all the estimations, this error can be added to the V_{output}.

Mitigation of error source: Curvature The main idea of curvature mitigation is introducing another same-aread constant-current-biased(biased by $I_{constant} = I_{PTAT}(T_r)$) BJT besides the origin one. According to 2.7, the difference between those two BJT becomes (2.37) [5].

$$\Delta V_{BE} = -(\eta - 1)V_t ln(\frac{T}{T_r}) - [-(\eta)V_t ln(\frac{T}{T_r})] = V_t ln(\frac{T}{T_r}) \tag{2.37}$$

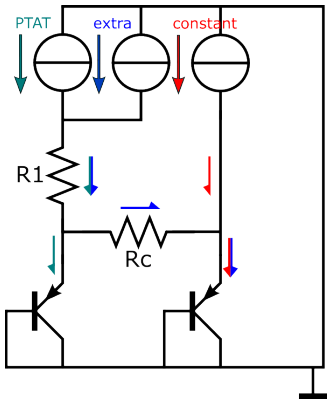

Figure 2.17: Ideal of canceling curvature error

As shown in Figure 2.17. The voltage difference between two BJTs can be transformed into a current by connecting a resistor R_c between those two BJTs. By forcing that current flow through R_1, the voltage difference is multiplied by $\eta - 1 = \frac{R_1}{R_c}$ and added to the output to cancel the $(\eta - 1)V_t ln(\frac{T}{T_r})$ curvature error from the PTAT-current-biased CTAT source. Ideally, the rest error behavior is in a third-order way because the second-order error (curvature error) is canceled. According to (2.9), the V_{ref} will also be lower because the V_G equals the bandgap of silicon at $T = 0K$.

One of the non-idealities is that, with the R_c connected between PTAT-current-biased and constant-current-biased BJT. The current created by R_c flows into the constant-current-biased BJT. This makes the V_{BE} of the constant-current-biased BJT no longer behave as expected. The V_{BE} of constant-current-biased BJT will behave more like the PTAT-current-biased one. Thus the $\Delta V_{BE} = V_{BE,PTAT} - V_{BE,constant}$ becomes less. This leads to only $\frac{1}{2}$ of expected ΔV_{BE} applied to R_c. One of the solutions to this is increasing the constant biasing current and using the same times of paralleled copies of the

constant-current-biased BJT. Thus the bias current for each copy would still be the same. But the extra current flows into each copy would be split.

The second method is to increase $\frac{R_1}{R_c}$. This requires a combination with the first solution because R_1 is not changeable due to its duty of creating a PTAT voltage drop. Only R_c can be decreased, but this would increase the current flow into the constant-current-biased BJT, which would decrease the ΔV_{BE} more. So copies are needed to split and attenuate this impact from extra current to the ΔV_{BE}.

So the best idea is to increase the bias current and the number of copies by n times and increase $\frac{R_1}{R_c}$ also by n times(note that if $\frac{R_1}{R_c}$ is increased too much than the constant-current-biased BJT will be biased in a worse case, which makes the error cancellation behaves worse). This would promise an almost unchanged ΔV_{BE}. But the coefficient for amplifying the ΔV_{BE} and using it to mitigate the curvature error is n times more. Normally $n = 2$ applied to this topology can decrease the curvature error to a negligible value.

Figure 2.18 shows an example in 16nm(which has a bigger curvature error than the normal case). The above five figures on the right show that the number of BJT copies and times of constant bias current keep increasing. The ΔV_{BE} gets closer to the ideal case. Thus the second-order error gets less and less, and the error at output becomes more like a third-order curve. But only doing this would take a lot of area to achieve an acceptable result. The two figures below show another way, increasing $\frac{R_1}{R_c}$ by the same ratio. Due to the redesign of R_c for the balanced output. The total extra current is more. But after being split by The copies, the bias current for each copy will be roughly the same as the mitigation design with the single copy BJT. But the coefficient for amplifying the ΔV_{BE} and using it to mitigate the curvature error is doubled. This leads to negligible error behavior in a third-order way. This also means increasing this aforementioned ratio n to 3 will lead to too much mitigation, resulting in a curve behaving in a second-order way combined with a positive coefficient for T^2, as shown in the lowest figure on the right. This over-mitigated curve is similar to the corrected curve in [2].

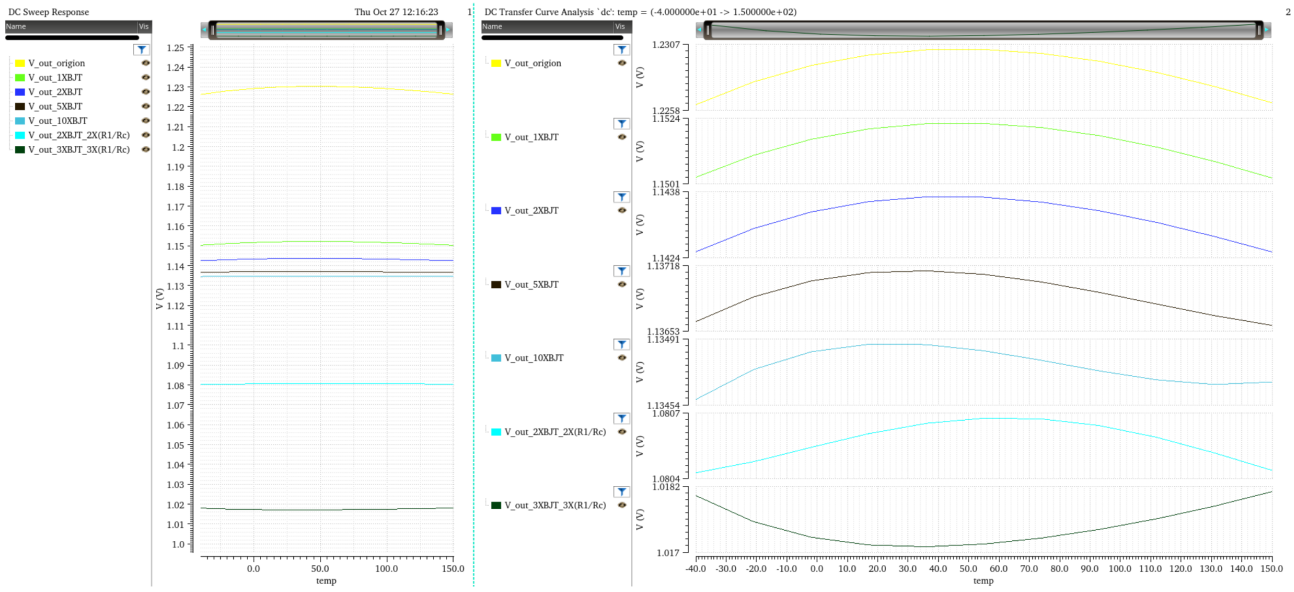

Figure 2.18: Curvature error under different design for mitigation

The other problem is the variation of the curvature correcting circuit during manufacturing which would make the constant bias current different from expected[2]. This can be taken as the correcting BJT being biased with a different bias current which leads to an extra $\frac{V_t ln(\frac{I_s}{I_{s_{correction}}} \frac{I_{c_{correction}}}{I_c})}{R_c}$ term into

I_{R_c}. This is in the same form as the error caused by saturation current variation, which can be removed by a one-temperature adjustment.

The curvature correction would always require extra components, which means more cost of Mitigation varies during manufacturing. In a previous design in NXP semiconductor using the same process, this error is accepted due to its small amount. But the aforementioned mitigation method will be required for the process with a bigger curvature error.

2.1.8 Error source: Noise

Introduction of error source: Noise Besides the error sources in the aspect of output voltage variation. Noise also brings errors into the output. Figure 2.19 shows the equivalent noise sources in a diode-connected BJT.

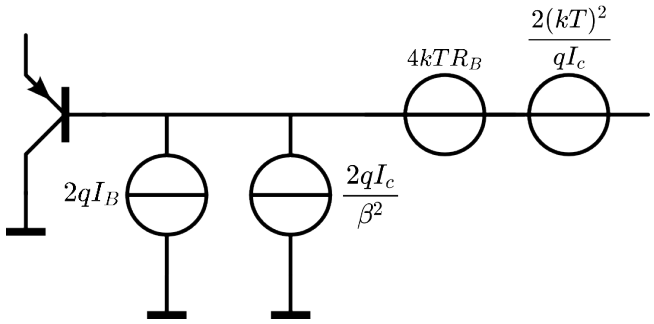

Figure 2.19: Noise sources of a diode connected PNP BJT

The flicker noise can be ignored if the excess-noise corner frequency is shallow. When the flicker noise is not negligible, a term $(1+\frac{f_l}{f})$ can be multiplied by the shot-noise sources. In which f_l is the excess-noise corner frequency[1]. This means a proper design regards the shot-noise sources can lead to an optimum design for the noise aspect.

Now it's time to discuss the noise sources. $\frac{2qI_c}{\beta^2}$ and $2\frac{(kT)^2}{qI_c}$ are the equivalent input noise sources of the collector shot noise[1]. $2qI_B$ represents the Base shot noise, and $4kTR_B$ is the thermal noise due to the base resistance.

The current noise sources can be ignored if A controller-made follower (Figure 2.14) is applied. This is because of the zero output impedance of the ideal followers. For the voltage noise, take $Ic = 1uA$ and $R_B = 400\Omega$ as examples. This leads to $4kTR_B << 2\frac{(kT)^2}{qI_c}$. Thus the noise from Base resistance can be ignored as well. The only noise left is $2\frac{(kT)^2}{qI_c}$

Mitigation of error source: Noise From (2.13), the noise power at the output can be delivered in (2.38).

$$S_{ref} = (a_1 + a_2)^2 \cdot \frac{(kT)^2}{qI_{c1}} + {a_1}^2 \cdot \frac{(kT)^2}{qI_{c2}} \tag{2.38}$$

Now insert (2.17), (2.18) into the equation for noise power at the output. (2.39) can be delivered. Note that I_{budget} is defined at $T = T_r$.

$$
\begin{aligned}
S_{ref} &= (a_1+a_2)^2\cdot\frac{(kT)^2}{qI_{c1}} + {a_1}^2\cdot\frac{(kT)^2}{qI_{c2}} \\
&= [\frac{V_{ref}-\frac{V_{ref}}{V_G}V_{BE1}(T_r)}{V_{BE1}(T_r)-V_{BE2}(T_r)}+\frac{V_{ref}}{V_G}]^2\cdot\frac{(kT)^2}{qI_{c1}} + [\frac{V_{ref}-\frac{V_{ref}}{V_G}V_{BE1}(T_r)}{V_{BE1}(T_r)-V_{BE2}(T_r)}]^2\cdot\frac{(kT)^2}{qI_{c2}} \\
&= [\frac{V_{ref}-\frac{V_{ref}}{V_G}\frac{kT_r}{q}ln(\frac{I_{budget}\cdot I_{c_{ratio}}}{I_{s1}})}{\frac{kT_r}{q}ln(\frac{I_{c_{ratio}}}{1-I_{c_{ratio}}}\cdot\frac{I_{s2}}{I_{s1}})}+\frac{V_{ref}}{V_G}]^2\cdot\frac{(kT)^2}{q}\cdot\frac{1}{I_{budget}(T)\cdot I_{cratio}} \\
&+ [\frac{V_{ref}-\frac{V_{ref}}{V_G}\frac{kT_r}{q}ln(\frac{I_{budget}\cdot I_{c_{ratio}}}{I_{s1}})}{\frac{kT_r}{q}ln(\frac{I_{c_{ratio}}}{1-I_{c_{ratio}}}\cdot\frac{I_{s2}}{I_{s1}})}]^2\cdot\frac{(kT)^2}{q}\cdot\frac{1}{I_{budget}(T)\cdot(1-I_{cratio})}
\end{aligned}
\tag{2.39}
$$

By varying the $I_{c_{ratio}}$ or I_budget, and keep all the other design variables to be constant($I_{s1} = 16e-8A, I_{s2} = I_{s1}\cdot n, T_r = 300K, I_{budget} = I_{c1}(T_r)+I_{c2}(T_r), V_{ref} = 0.2V$). Figure 2.20 is shows an example of the noise power at $T = 300K$. These parameters are the same as the ones used in [1], which will be used to compare the difference between the PTAT+CTAT model and the CTAT-CTAT model later.

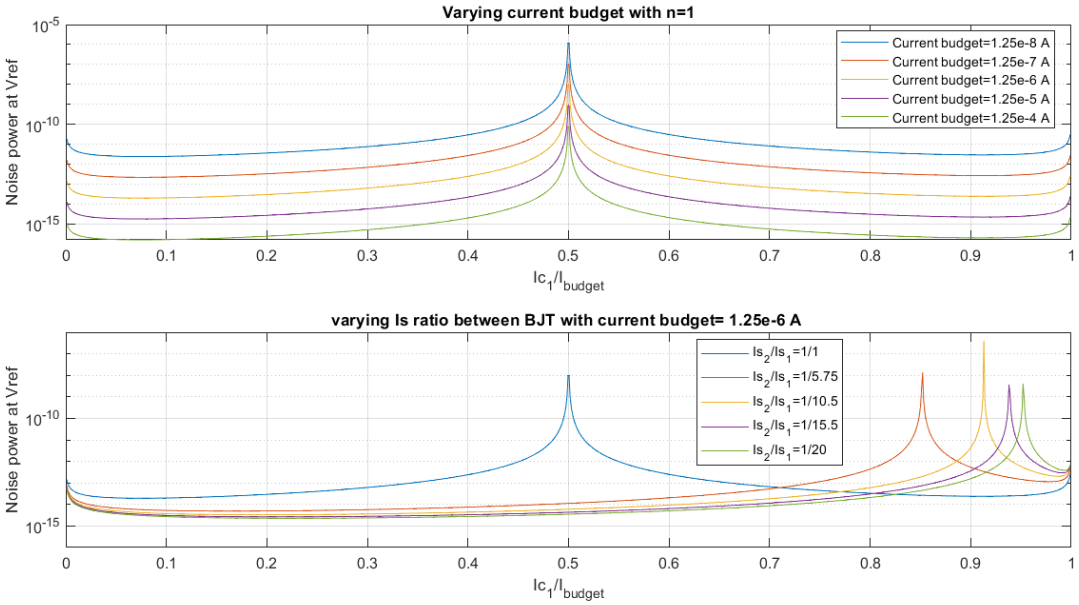

Figure 2.20: An example of the noise power at output with different $I_{c_{ratio}}$ under different I_{budget} or $\frac{I_{s2}}{I_{s1}}$ base on PTAT+CTAT model

The general noise level is defined by the I_{budget}, this is because both two terms in the equation has a $\frac{1}{I_{budget}(T)}$ and this $I_{budget}(T)$ is defined by $(0A, 0K)$ and $I_{budget}(T_r) = I_{budget}$. So for all temperatures, the power will be multiplied to $\frac{T}{T_r}$. The first figure shows that the lowest noise is limited by the I_{budget}, and for different $I_{budget}(T)$, they are just the same curve shifted in Y-axis.

In second figure, $I_{c_{ratio}}$ varies under different $\frac{I_{s2}}{I_{s1}}$. The noise explodes when $I_{c_{ratio}}$ is close to 0 or 1. This is because at those two cases, $\frac{1}{I_{c1}} = \infty$ or $\frac{1}{I_{c2}} = \infty$. It also explodes when $\frac{I_{c_{ratio}}}{1-I_{c_{ratio}}} = \frac{I_{s1}}{I_{s2}}$. This is because, in that case, $V_{BE1} = V_{BE2}$, which requires an extremely large a_1 to get a balanced

V_{output}. This also amplifies the noise as an extremely large value at the output. Thus there will be a lowest-noise $I_{c_{ratio}}$ for each I_{budget}. And for different I_{budget}, the optimum $I_{c_{ratio}}$ is the same.

The optimum point can be found by taking $\frac{d(S_{ref})}{d(I_{c_{ratio}})} = 0$

With the same parameters, taking $\frac{I_{s1}}{I_{s2}} = 10$ and $\frac{I_{s1}}{I_{s2}} = 0.1$ as an example. Plot the PTAT+CTAT model and CTAT-CTAT model (analyzed in [1]) in the same sketch, as shown in Figure 2.21. It can be seen that PTAT+CTAT has a lower optimum point in $\frac{I_{s1}}{I_{s2}} = 10$ case. And CTAT-CTAT behavior better in $\frac{I_{s1}}{I_{s2}} = 0.1$ case. This is because the PTAT+CTAT model is not a symmetric model. One of the noise sources has an extra part in its coefficient, which leads to a better optimum point in the PTAT+CTAT model.

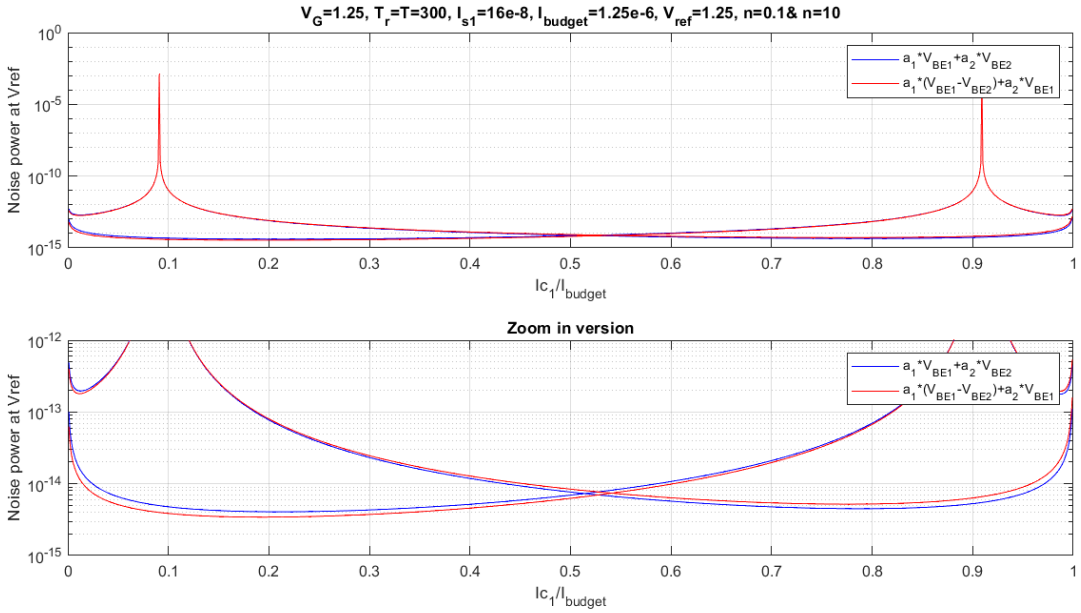

Figure 2.21: A comparation between PTAT+CTAT model and CTAT-CTAT model

Since noise is not the primary error source in this project, the conclusion for this error source will be a suggestion when choosing design parameters.

2.1.9 Conclusion for BJT errors

With the discussion of all the error sources, now it's time to have a summary of them. This could generate a list of design parameters and how the designer should value them.

Saturation current This is an error that does not exist in the nominal case. It changes the CTAT source(V_{BE}) both in process and mismatch way. This is an error that can be fully removed with a one-temperature adjustment. While designing, a smaller $|a_1|$ coefficient can attenuate the mismatch variation leading to error. And a bigger $\frac{I_{s1}}{I_{s2}}$ or bigger $\frac{I_{c2}}{I_{c1}}$ leads to lower $|a_1|$.

Current gain This error exists in the nominal case if the BJT is diode connected by connecting its Base and Collector with a single wire. In the nominal case, the error caused by the current gain can be ignored, but the process and mismatch variation of the current gain leads to a non-negligible part that cannot be removed. To trim this error a two-temperature adjustment is required. A bigger

Table 2: Conclusion for design parameters of BJT

Design parameters	**Max limiting aspects**	**Min limiting aspects**	**Typical value**
I_{s2}	available area for BJT2	minimum area for BJT	min
I_{s1}	available area for BJT 1	error amplified by a_1	$> 10 \cdot I_{s2}$
I_{budget} at T_r	power consumption Base resistance caused error	noise contribution of both BJT	$1\mu A$
$\frac{I_{c1}}{I_{budget}}$	noise contribution of BJT2	noise contribution of BJT1 error amplified by a_1	0.2 for $\frac{I_{s1}}{I_{s2}} = 10$
T_r	room temperature	room temperature	$27°C$

β can optimize the behavior but costs a lot of area. A follower made by controller applied to BJT can prevent β from participating in the output, which replaces the old errors with the errors from the controller. A lower $|a_1|$ can attenuate the error caused by the current gain mismatch variation(if do not apply follower) and the offset of the controller.

Base resistance This is an error that exists in the nominal case. It can be made negligible by using a low Base current. In the given process, $1\mu A$ applied to the Collector current is low enough.

Curvature This is an error that exists in the nominal case. It is an additional variance that has nothing to do with the design. This can be mitigated by introducing another curvature error but it requires more components. Considering the $\pm 0.5\%$ target in this project, straightly accepting this error can also be a solution.

Noise The noise from the BJT can be attenuated by supplying more current to the Collector or choosing an optimum $\frac{I_{c1}}{I_{c2}}$. The biasing current limits the performance of the optimum point. $\frac{I_{c1}}{I_{c2}}$ enables the design to be at the optimum point. The designer should avoid $I_{c1} \approx 0A$, $I_{c2} \approx 0A$, and $\frac{I_{c1}}{I_{c2}} = \frac{I_{s1}}{I_{s2}}$.

Summary The aforementioned error sources deliver a list of design parameters and their maximum and minimum limiting aspects, as shown in table 2.

2.1.10 A Design Example regards errors refer to BJT

To check if the aforementioned errors in the process and mismatch variation are valid, a circuit build-up by ideal resistors and controller with a huge gain, as shown in Figure 2.22 is used. This circuit is set up in Cadence for running the Monte Carlo test to check if the error estimation is correct. Since the simulation tool can not simulate the error sources separately, the error result is available for the combination of the errors.

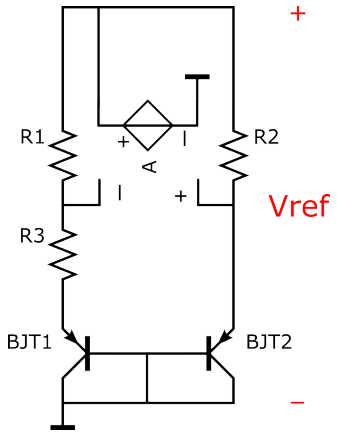

Figure 2.22: Test circuit used in Cadence

As mentioned in (2.13), a pair of a_1 and a_2 is needed. In this circuit, a_2 is assigned to be 1 to make the circuit simple, a_1 is realized by $-\frac{R_1+R_3}{R_3}$. According to table 2, following design parameters are chosen:

BJT2 is designed to have an Emitter area of $2\mu m \cdot 2\mu m$, $\frac{I_{s1}}{I_{s2}} = 10$, $I_{c_{ratio}} = \frac{I_{c1]}}{I_{c2}} = 0.2$, and $I_{budget} = 1\mu A$. BJT1 is a 10-fingers-paralleled-connected version of BJT2.

The first thing to check is the V_{BE} at the nominal case, as shown in Figure 2.23. Note that this model is the V_{BE} considering the current gain's impact on the Collector current. And The models are correct enough to be used for design.

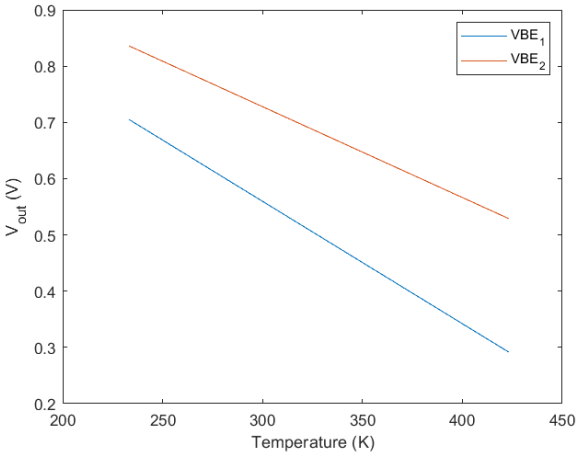

Figure 2.23: V_{BE} model of the design example

Now, with knowing $a_2 = 1$ and inputting the V_{BE} model into (2.13). By making sure $V_{output}(233K) = V_{output}(423K)$, the a_1 specially designed for the nominal case is calculated. The calculated V_{output} at different temperatures is shown in Figure 2.24.

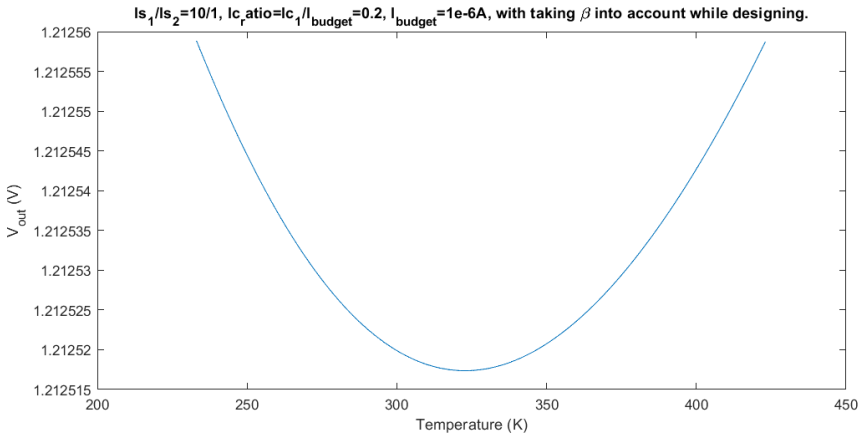

Figure 2.24: Calculated $V_{out\,put}$ in nominal case

By applying the calculated R_1, R_2, R_3 into the circuit in Cadence, the $V_{out\,put}$ is shown in Figure 2.25.Due to the mistakes made while taking parameters from the model, the result is not the optimum case. The optimum case is determined by slightly adjusting the resistors and shown in Figure 2.26.

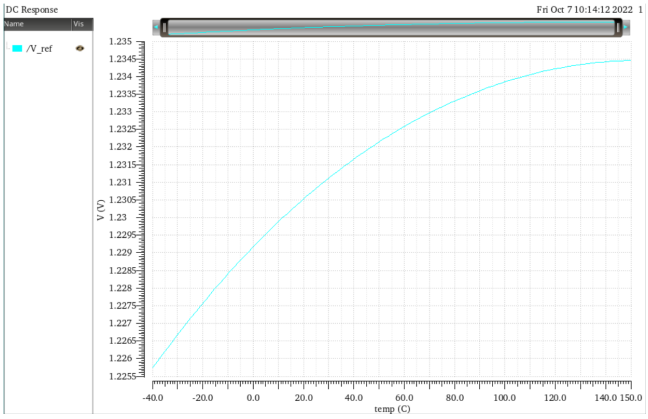

Figure 2.25: Calculated result behaves in Cadence

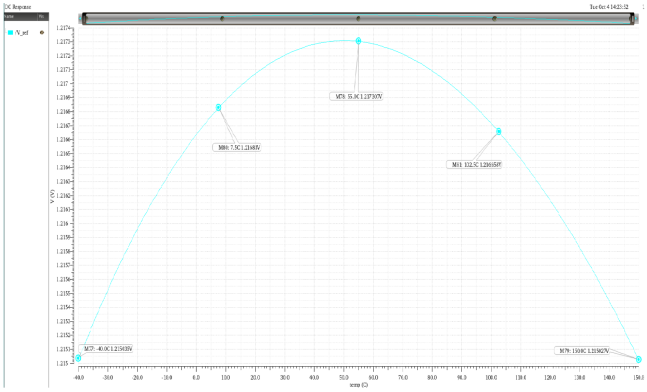

Figure 2.26: Adjusted result behaves in Cadence

Note that principle models did not consider the curvature errors. With the optimum designed a_1 and a very low bias current, the only error left in Figure 2.26 is the curvature error, which is around $2.5mV$ in the nominal case.

This means that the parameter modeling of the principle model can represent the models from cadence in a valid way. With the way of modeling variations in section 2.1.4 and 2.1.5. A plot of the worst cases is delivered in Figure 2.27. This result can be the estimated Monte Carlo result based on the principle model. The variation of each case in the percent of the nominal case is also plotted. Note that the Monte Carlo result from Cadence should have a slightly lower variance. This is because the current gain and saturation current variations are not independent. Thus, some extreme occasions do not exist in the required variation range (5σ).

To check if the estimated result is valid, a Monte Carlo simulation is done in Cadence. Due to the simulation time of the Monte Carlo test, the output of 5 temperatures($-40°C$,$7.5°C$,$55°C$,$102.5°C$,$150°C$) is picked up to form the result of variations, as shown in Figure 2.28, and the variation in the percent of the nominal case is also plotted. The figure below shows that the curvature error is roughly the same for variations. Because by making the difference between nominal and variations, this curvature error can be removed. This is why the V_{output} performs a curvature error, but the error does not. Therefore, the result from Cadence also shows that the error estimation is valid.

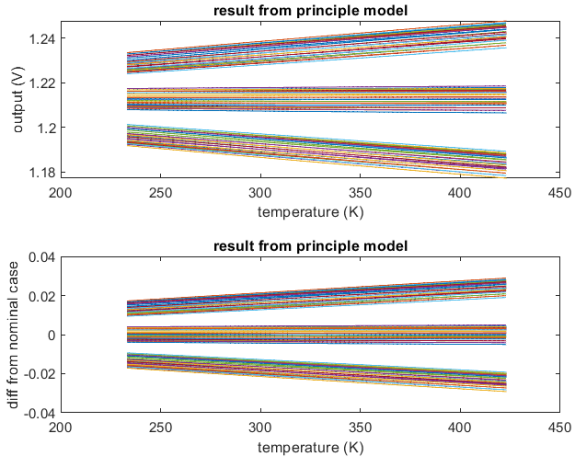

Figure 2.27: Estimated Monte Carlo result in Principle model

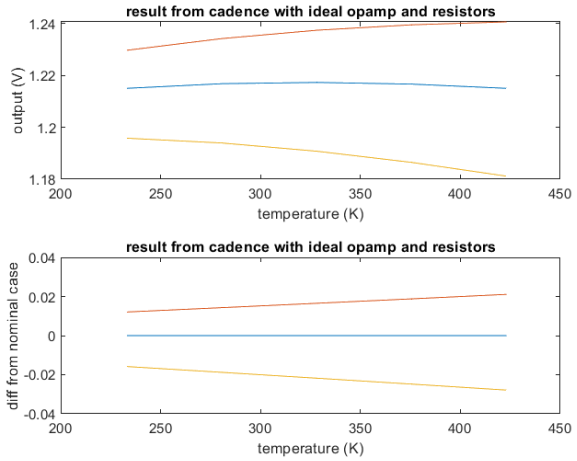

Figure 2.28: Monte Carlo simulation result in Cadence

Now at the end of the design example, the table shows how much each error source contributes

Table 3: Error contributions in the design example

Error sources	**Typical value**	**Error contribution**	**Cancellable**
Saturation current spread	process: $(-55\%, +122\%)$ mismatch: $(-0.8\%, +0.8\%)$	$\pm 2.5\%$ before adjustment 0% after adjustment(ideal)	fully adjust by one temperature
Current gain spread	process: $(-25\%, +47\%)$ mismatch: $(-2\%, +2\%)$	$\pm 0.4\%$ before adjustment $\pm 0.3\%$ after adjustment	partly adjust by two temperature
Base spread resistance	400Ω	$\pm 0.06\%$	no need
Curvature error	$2.5mV$	$\pm 0.1\%$	no need

to the output in the given design BJTs when no adjustment is applied. In addition, the left error after trimming will be shown as well. This helps the designer with deciding to focus on which error to mitigate.

From the list, an error mitigation strategy can be made. The critical error in BJT in the given process is the saturation current spread. Thus it will be better to let the adjustment on a_1 entirely focus on canceling that. This requires a design that transforms the current gain spread into another error aspect, as mentioned in section 2.1.5. With two followers whose offset is within $\pm 5mV$, the error caused by the current gain spread can be replaced by an offset-caused error range in $\pm 0.2\%$. In some processes, the current gain is high enough to provide a non-adjusted error less than $\pm 0.2\%$. Then there is no need to implement the follower. This also enables the a_1 to entirely focus on canceling the saturation current spread. Then the after-adjusted saturation-current-leading error will be only limited by the adjustment resolution of a_1. By accepting the Base spread resistance and curvature error, there is still some area left. Since the dominant error among the left errors after adjusting a_1 is the offset of the followers, it is still possible to sacrifice some area for a smaller $|a_1|$.

The analysis progress shown in this section can be made for the PNP or NPN bipolar junction transistors in other processes. This leads to a good enough design of BJT pairs that provides the CTAT source (V_{BE}). The analysis progress of the rest part will be in the following sections.

2.2 Concept circuit level

With the conclusion from the last section, it's time to use a concept circuit to realize the principle model with a concept circuit. With the concept circuit, how the error sources from other components and their requirements will be worked out. Which will be realized with schematic-level design in detail.

This section will start with listing the mathematical relationships and how they will be realized. Then a topology based on that will be proposed, and so does the requirements for each component. For a clear explanation, examples in process C16FFC will be used.

2.2.1 Generating of the CTAT source

The first step is to generate two different V_{BE}, as shown in Figure 2.29. This can be realized by two BJTs biased with different bias current densities, no matter how much current or the saturation current is, as long as the bias current density is different, different V_{BE} can be generated.

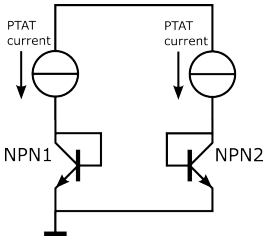

Figure 2.29: Two different bias current density leads to two different V_{BE}

2.2.2 Generating of the PTAT biasing source

As shown in Figure 2.30. With the difference in two V_{BE}, try to force them on two sides of resistor R_3 with a nullor; a PTAT current will be generated and supplied by the nullor. And by applying resistors R_1 and R_2, the current injected into two BJTs will be defined by them.

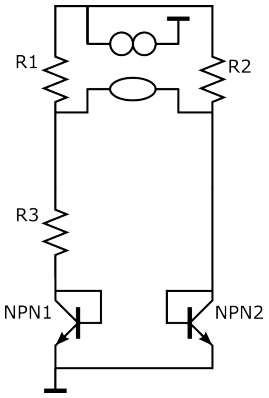

Figure 2.30: Generate PTAT bias current with two CTAT sources

Another popular way to realize the current relationship is the current mirror, as shown in Figure 2.31. This structure leads to a lower voltage gain requirement for the controller used to realize the nullor. To compare the difference between using resistor and current mirror to split the current, the V_{out} needs to be located first. They are shown in Figure 2.32 and Figure 2.31. The detail of making the V_{out} being constant is shown in section 2.2.3

The transfer from the input offset to the output offset in Figure 2.32 is shown in (2.40).

$$\begin{aligned}\frac{V_{out}}{V_{in}} &= 1/\left(\frac{\frac{1}{gm_{NPN2}}}{R_2+\frac{1}{gm_{NPN2}}} - \frac{R_3+\frac{1}{gm_{NPN1}}}{R_1+R_3+\frac{1}{gm_{NPN1}}}\right) \\ &\approx -\frac{R_1+R_3}{R_3} \\ &\approx -5\end{aligned} \tag{2.40}$$

The transfer from the input offset to the output offset in Figure 2.31 is shown in (2.41).

$$\begin{aligned}\frac{V_{out}}{V_{in}} &= \frac{1}{gm_{p1}\cdot R_3}\cdot gm_{p2}\cdot(\frac{1}{gm_{NPN2}}+R_2)\\ &\approx -\frac{R_2}{R_3}\\ &\approx -4\end{aligned} \tag{2.41}$$

The calculation shows that with the current mirror splitting the current, the output mismatch can be slightly less, but it's still not ignorable. This means it would require similar mismatch attenuation methods in future design steps. Then it's time to discuss if it's more effective to realize current splitting with current mirrors.

The cost of realizing the same accuracy in current splitting realized by resistors with the current mirror is 1700 times more in area in the given process 16FFC. And in 16FFC, the designer can only increase the effective Length of the Finfet by connecting multiple Finfets in series and connecting the gate of them with each other. Thus the V_{ds} of the extra Finfets may lead to risk in biasing the Finfet work as the current source in the triode region. Besides, the current mirror's V_{ds} difference also leads to systematic error. Thus considering the cost of area, resistors lead to a more promising solution in the given process 16FFC.

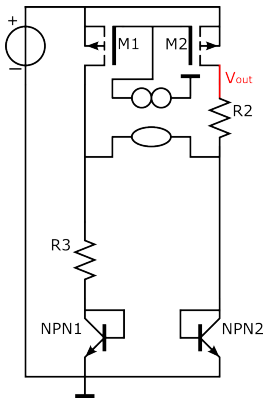

Figure 2.31: Generate PTAT bias current with two CTAT sources and apply current mirror

2.2.3 Forming constant output with PTAT and CTAT source

Since in Figure 2.30, the current flows through R_1 and R_3 are the same. Mark the output node of the nullor as the V_{out}, as shown in Figure 2.32. This means that V_{out} can be expressed in (2.42).

$$V_{out} = \frac{R_1}{R_3}\cdot(V_{BE2}-V_{BE1})+V_{BE1} \tag{2.42}$$

$$V_{BE2}-V_{BE1} = V_t\cdot ln(\frac{I_{c2}}{I_{s2}}\cdot\frac{I_{s1}}{I_{c1}}) \tag{2.43}$$

$$\frac{I_{c1}}{I_{c2}} = \frac{R_2}{R_1} \tag{2.44}$$

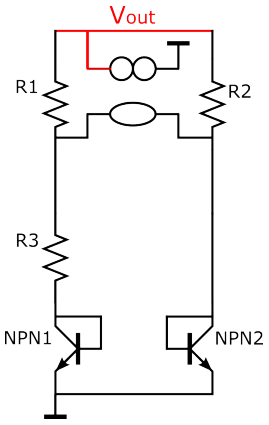

Figure 2.32: Forming constant output with PTAT and CTAT source

This points out that by properly designing the resistance of R_1, R_2, and R_3. The first-order part from the CTAT and PTAT can be canceled by each other. According to (2.15), when the first-order temperature-dependent part is removed, this leads to a $V_{out} = V_G$.

Note that a bigger V_{BE} difference leads to less $\frac{R_1}{R_3}$ required. And the input offset of the controller used to realize the nullor in the future will also be amplified by the close loop gain$= \frac{R_1}{R_3}$. Because the current will be defined by $\frac{V_{BE2}-V_{BE1}-V_{os}}{R_3}$. Thus in the design step, difference between $\frac{I_{c1}}{I_{s1}}$ and $\frac{I_{c2}}{I_{s2}}$ needs to be designed as big as possible.

2.2.4 Canceling the curvature error

The solution to removing the first-order error is realized in the last section. Now it's time to discuss how to remove the second-order error from the V_{BE}. According to section 2.1.7. As shown in 2.33, if properly design and realize $\frac{R_1}{R_{c1}} = \frac{R_2}{R_{c2}} = \eta - 1$. Ideally, the second order error in V_{BE} can be canceled and lead to V_{out} in(2.45.)

$$V_{out} = \frac{R_1}{R_3} \cdot (V_{BE2} - V_{BE1}) + V_{BE1} + (\eta - 1) \cdot V_t ln(\frac{T}{T_r}) \tag{2.45}$$

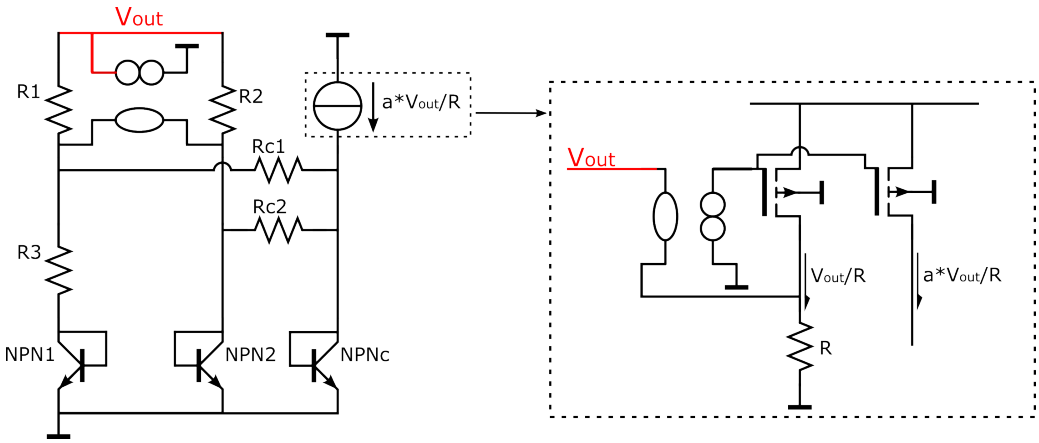

Figure 2.33: Canceling the curvature error

Another solution is to use Howland current pump to generate such current, as shown in Figure 2.34. But Howland current pump has positive and negative feedback, which would lead to extra effort

in design required to promise stability. And since this output stage current mirror is not the dominating offset error source(the dominating one is the input offset of the core controller). It is better to start with the current mirror, and if the cost of area for making the mismatch of the current mirror to be low enough is too much, go back to the Howland pump.

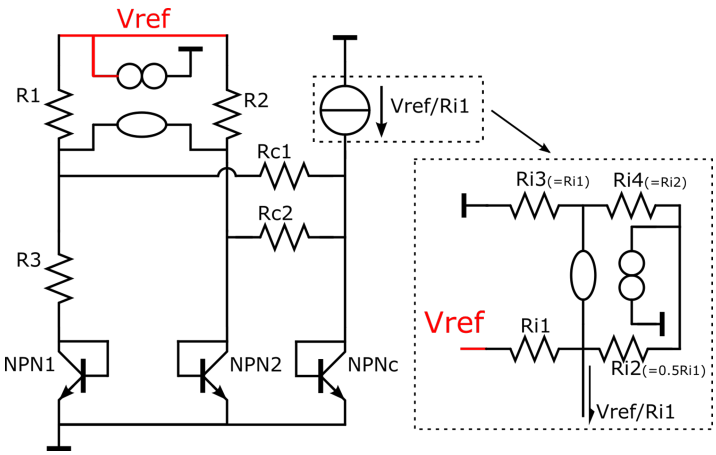

Figure 2.34: Canceling the curvature error with Howland current pump

2.2.5 Implementing of trimming

Since R_1 and R_2 are connected with the same voltage on both sides. This means that part of them can be split from them and replaced by another new resistor. The coefficient for amplifying the PTAT voltage can be redefined by trimming that resistor, as shown in Figure 2.35.

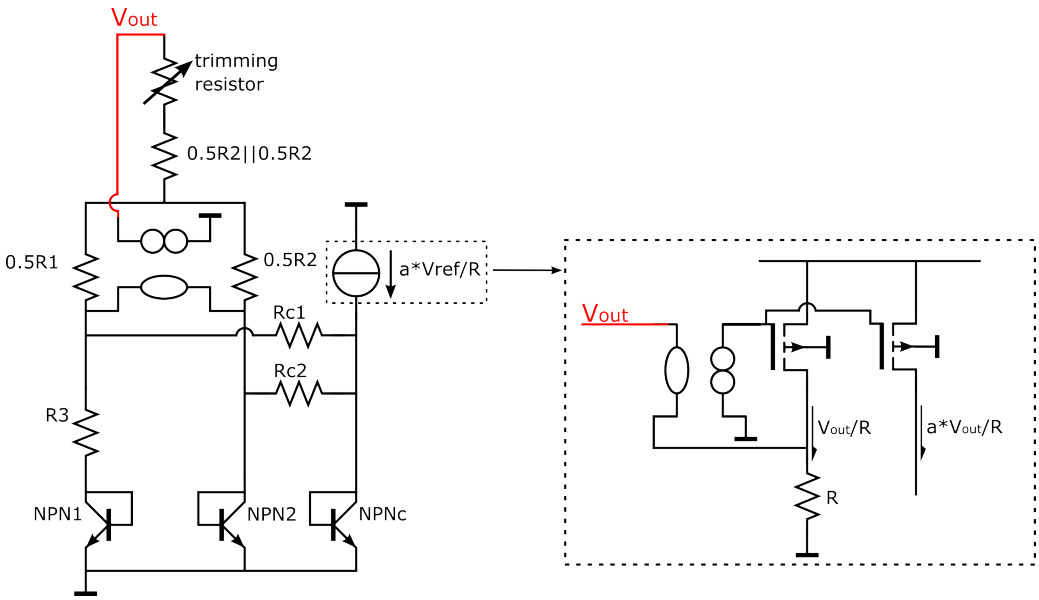

Figure 2.35: Implementing of trimming

2.2.6 Replace nullor with controller

Before moving to the schematic level, the nullors need to be realized by the controllers, and the polarities of the controllers need to be assigned, as shown in Figure 2.36. Besides that, a low pass

filter is added to the output node in case of noise cancellation. Since the load connected to this bandgap circuit will be only capacitive, an RC filter would be a good solution to this low-pass filter.

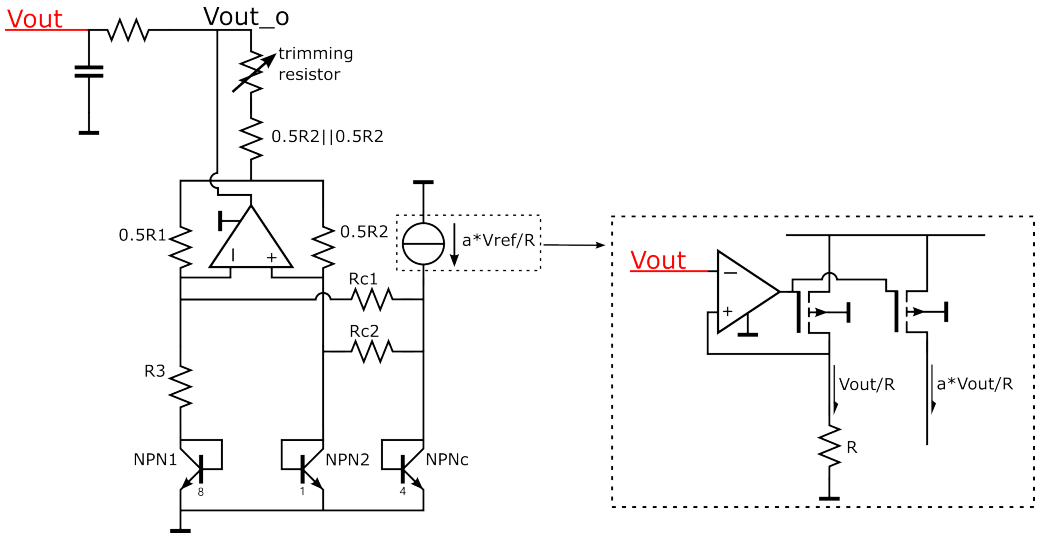

Figure 2.36: Replacing the nullors with controllers and add low pass filter

2.2.7 Gain from input offset to output offset

As mentioned in previous sections, the key to accuracy is the part of error that cannot be trimmed, which is the offset. And the main offset source in V_{out} is the input offset of the core controller. As mentioned in (2.40), this gain will be roughly 5. But with the curvature correction part, this gain increase to $\frac{R_1}{R_3}+1+\frac{R_1}{R_{c1}} \approx 10$. In the given process, it is quite common for the input offset of A controller to reach $\pm 10mV$. This leads to $\pm 100mV$ which is too much. Chopping can be applied to the controller to minimize the input offset.

2.2.8 Conclusion for concept circuit level

The key to the whole circuit's performance is the post-trim accuracy of V_{out}, which is dominated by the core controller's input offset. So the next step is to design the controllers with as less input offset as possible, which will be discussed in the next section. The detailed design decisions of the concept-level circuit behavior will be shown in section 3.2.

2.3 Schematic level

In the schematic level design, the transistor level topology for the controllers and the trimming resistor will be generated, and the contribution from each component of the topology to the output offset will be discussed as well. The core controller will be discussed first and then the controller for the constant current for curvature correction called as second one.

2.3.1 Design for effective area

Bandgap reference circuits are the type of circuit that normally does not take too much power and area in the chips. In in-vehicle networks, the power budget is not strict, so designers prefer them to be as small as possible and within an acceptable power. And from the aspect of circuit behavior, the aspect

which dominates the area is the mismatch and sometimes the flicker noise. Thus for an effective area design, it is better to realize the circuit behavior first and then increase the area of the transistors to meet the mismatch and flicker noise requirements. This shall promise a design version that meets the requirements and uses as little area as possible.

2.3.2 Core controller: Input stage

Ideally, the input of the core controller is wished to be made exactly the same by the feedback, an input pair would be a good solution. The first thing which can already be told from the concept circuit is the common mode input voltage of the input pair of the core controller. The result of design decisions from section 3.2 shows that the common input level will be in $520mV \sim 850mV$ in $-40°C \sim 150°C$. Considering the given VDD is $1.8V$, there is no need to do pseudo differential input pair. So a pair of PMOS as the input pair would be good enough, and a PMOS current source for protecting the bias condition is also added, as shown in Figure 2.37.

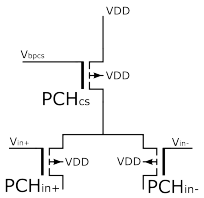

Figure 2.37: Input pair of the core controller

2.3.3 Core controller: Output stage

An output stage work as a sourcing current source is needed to supply the current biasing of the BJTs, as shown in Figure 2.36. This can be realized by a PMOS working as current source as the output stage of the controller, as shown in Figure 2.38.

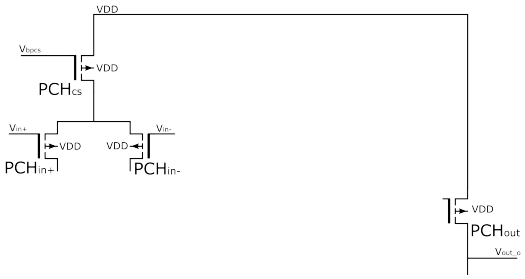

Figure 2.38: Output stage of the core controller

2.3.4 Core controller: Current mirror

Now adding current mirrors to connect and take advantage of the gain of both differential sides, As shown in Figure 2.39

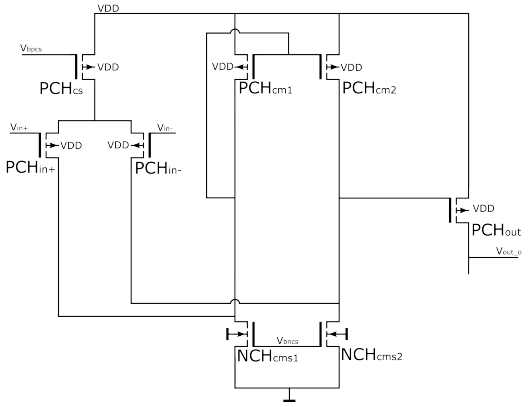

Figure 2.39: Adding current mirrors to the core controller

2.3.5 Core controller: Cascode stage

Considering the given VDD is high enough($1.8V$), and the threshold voltage of the given Finfet is around $400mV$. The VDD allows more than 2 Finfet connected in series. Besides that, the controller in the Bandgap reference circuit does not need good linearity regards the swing range. Thus Finfet work as common gate stages can be added to the circuit for more output impedance, As shown in Figure 2.40

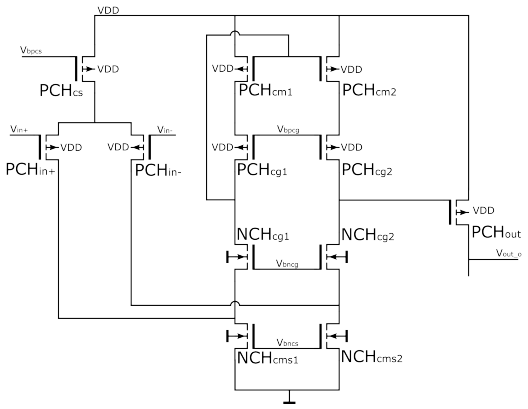

Figure 2.40: Adding common gate stages to the core controller

2.3.6 Core controller: Voltage gain and dominant pole

From the topology, the voltage gain delivered by the core controller can be calculated, as shown in (2.46)

$$A_{DC} = 2 \cdot gm_{in} \cdot [(gm_{cgp} \cdot r_{d,cgp}) \cdot r_{d,cmp} || (gm_{cgn} \cdot r_{d,cgn}) \cdot (r_{d,cmn} || r_{d,in})] \cdot gm_{out} \cdot R_{load} \tag{2.46}$$

For making the dominant pole fully under the designer's control and promising a good power supply rejection ability. Normally the pole established by the R_{out} of the cascode stage and the gate capacitor of the output stage would be made as the dominant pole. This is because the components in the controller are something that is fully under control. And aforementioned components are the biggest resistance and capacitance in the controller. The dominant pole will locate at (2.47).

$$pole = \frac{1}{2\pi RC} = \frac{1}{2\pi \cdot [(gm_{cgp} \cdot r_{d,cgp}) \cdot r_{d,cmp} || (gm_{cgn} \cdot r_{d,cgn}) \cdot (r_{d,cmn} || r_{d,in})] \cdot C_{g,out}} \tag{2.47}$$

2.3.7 Current budget and gm/ID design method

After finishing the topology design, the next step is assigning each branch's current budget. This is because power and area are the most important budget in nowadays design budget of analog IC design. And with fixing the current budget of each branch, the designer can also find a starting point for designing based on that.

Taking the input pair of the controller(PCH_{in+},PCH_{in-}) as an example. The input pair directly influence two controller aspects: Noise and gain. The input pairs' thermal noise($\propto \frac{1}{gm}$) and flicker noise($\propto \frac{1}{W \cdot L}$) both contribute to the output referred noise. The input pairs' transconductance (gm) contributes to the controller's gain. This means that by fixing the bias condition(fixing gm), no matter how much area($W \cdot L$) is spent on the input pairs, the thermal noise and the gain will be the same. By establishing a gm/ID model, the designer can find a starting point with a good enough transconductance(gm).

With a nullor helping with biasing the Finfet, the biasing condition of the Finfet can be dominated by the biasing current; the nullor supplies the needed gate voltage, as shown in Figure 2.41.

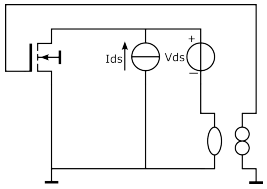

Figure 2.41: Circuit for establish gm/ID model

In Finfet process, it is always good to design the Finfet to have the same number of Fins and only design the fingers, this leads to more interarea saved when designing the layout. Thus by inserting the same amount of current into different Fingers of Finfet, a model expressing the relationship between gm and $ID/Finger$ is established. Figure 2.42 shows gm and noise of an NMOS biased with $5uA$ with $L = 240nm$, $Fin = 20$, and different number of Fingers.

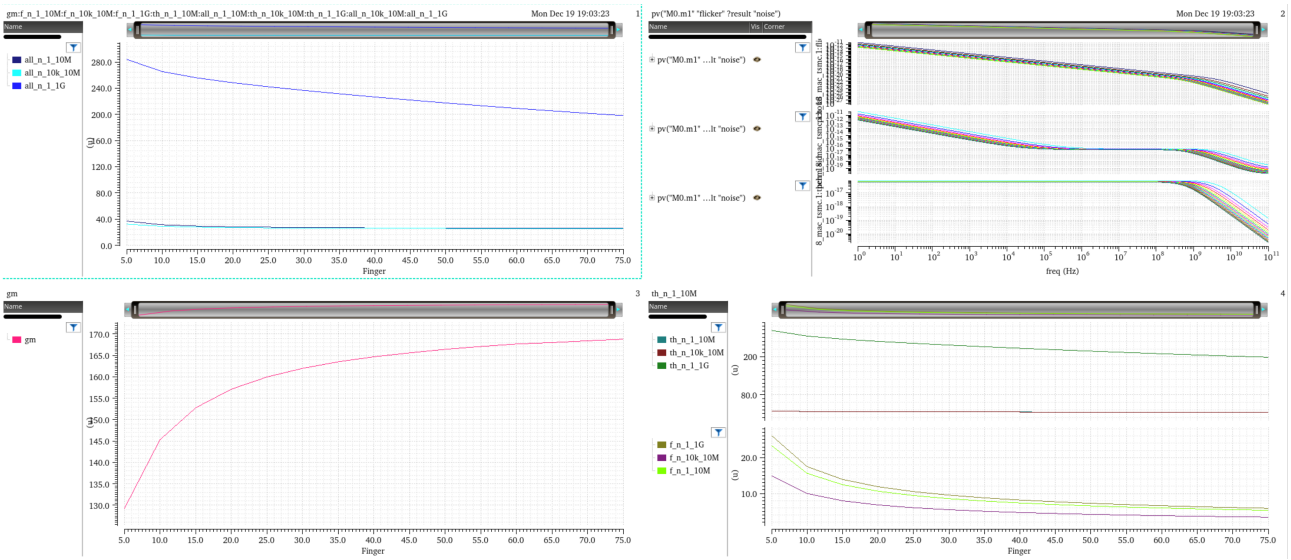

Figure 2.42: gm and noise of NMOS biased with $5uA$ with $L = 240nm$, $Fin = 20$, and different number of Fingers

As shown in the bottom left figure, the Finfet is biased in weak inversion with increasing the number of fingers. And the gm increases in a saturation way while the number of fingers increases. This means that the designer can find a starting point that the number of Fingers leads to a budget effective gm. This can also tell the designer that if the integrated thermal noise is not reasonable, then more current budget must be spent on biasing the Finfet. If the flicker noise is too much, then the designer can increase the W and L by the same ratio to promise the same bias condition. The designer can also only increase W. This would decrease the BW of the Finfet and slightly increase the gm(in the right top figure, the thermal noise spectral density of more Finger does not decrease significantly, because the gm/ID starts saturating when the designer keeps increasing the Finger after passing the effective number of Fingers.). The top left figure shows the total integrated noise in $1Hz$ to $1GHz$ also saturated at some number of Finger, this is because the flicker noise is low enough to be ignored. And the integrated thermal and flicker noise in different BW is separately shown in the bottom right figure. This model can be used for the Finfet works as different roles in the topology.

2.3.8 Core controller: Bias circuit

After planning the current budget and the area for the Finfet in the controller topology, the bias circuit for generating such a bias current needs to be designed. Figure 2.43 shows two pairs of current mirrors establishing a self-bias circuit and another two branches generating gate voltages for controlling current sources. (2.48) shows how the current in the self-bias current mirrors is defined, $I_{self-bias}$ is the current flows in the resistor R_{bias}.

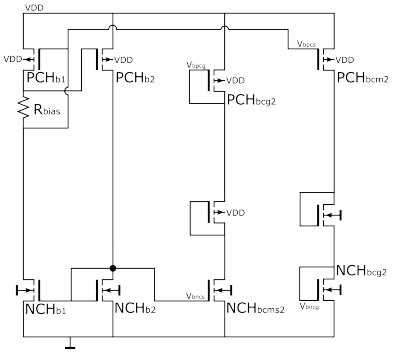

Figure 2.43: Bias circuit

$$I_{self-bias} = \frac{2}{\mu_p C_{ox}(\frac{W}{L})_{PCH_{b1}}} \cdot \frac{1}{R_{bias}^2} \cdot (1 - \frac{1}{\sqrt{\frac{(\frac{W}{L})_{PCH_{b2}}}{(\frac{W}{L})_{PCH_{b1}}}}}) \tag{2.48}$$

2.3.9 Core controller: Bias Start-up circuit

But there is another occasion that also meets $I_{ds,NCH_{b1}} = I_{ds,NCH_{b2}}$, which is $I_{ds,NCH_{b1}} = I_{ds,NCH_{b2}} = 0$. This can be avoided by adding a start circuit realized by a diode-connected Finfet, as shown in Figure 2.44. If both the PMOS current mirror and NMOS current mirror are closed, the diode-connected Finfet will try to pull the gate voltage of the PMOS current source down and avoid 0 current occasions occurring. Note that in practice, the designer may have to connect several diode-connected Finfet in series in case of the difference between the gate voltage of the PMOS current mirror and NMOS current mirror at the steady state are bigger than the threshold voltage(V_{th}) of one Finfet. The designer must fulfill the relationship in (2.49) to let such a start circuit work properly. The first formula is for letting the start circuit start working at the initial occasion. The second formula is for letting the start circuit stops working at the steady state. For high sigma design, the designer may need to increase the area of the whole bias circuit to ensure the threshold voltage variation is in an acceptable range.

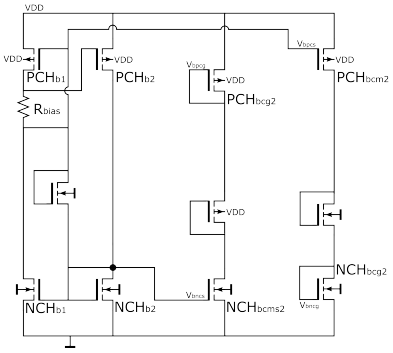

Figure 2.44: adding start circuit to the Bias circuit

$$\begin{cases} V_{th,NCH} + V_{th,start} + |V_{th,PCH}| < VDD \\ V_{GS,NCH} + V_{th,start} + |V_{GS,PCH}| > VDD \end{cases} \tag{2.49}$$

2.3.10 Core controller: output stage Start-up circuit

This kind of 0 current occasions could be possible at the top level, as in Figure 2.36. If the output current of the output stage of the core controller is almost 0(for example *pA* level), then the input pair of the core controller can also get an equal input due to the extremely low voltage drop on the resistor and almost no difference in V_{BE}. A start circuit that always pulls the gate voltage of the output stage down when the V_{BE} is not high enough can be added to the circuit to avoid such occasions, as shown in Figure 2.45.

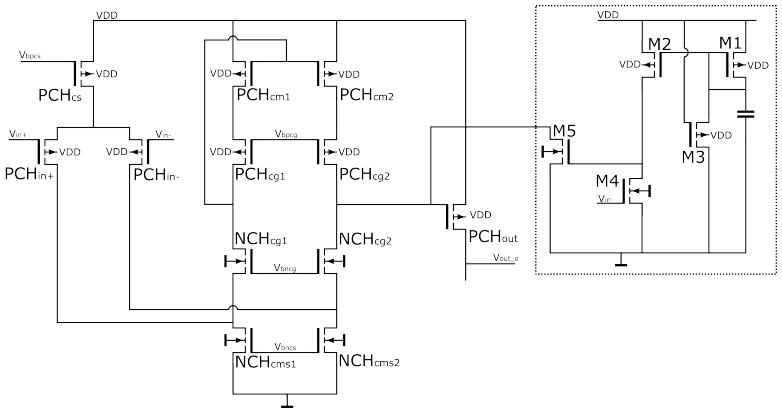

Figure 2.45: adding start circuit to the output stage of core controller

During power on, the gate capacitor of M5 will be charged, and M4 will be closed. This would force the gate voltage of the output stage to be pulled down to GND. When the input pair's input voltage (V_{BE}) is high enough to open M4, the gate charge of M5 will be released. The feedback will help the output stage stay at its steady state. During the pulling down of the gate voltage of the output stage, there may be spike in the current sourcing by the output stage, because the gate voltage is close to GND. The designer can add some diode-connected Finfet as charge protector to avoid $V_{g,output} = GND$, as shown in Figure 2.46.

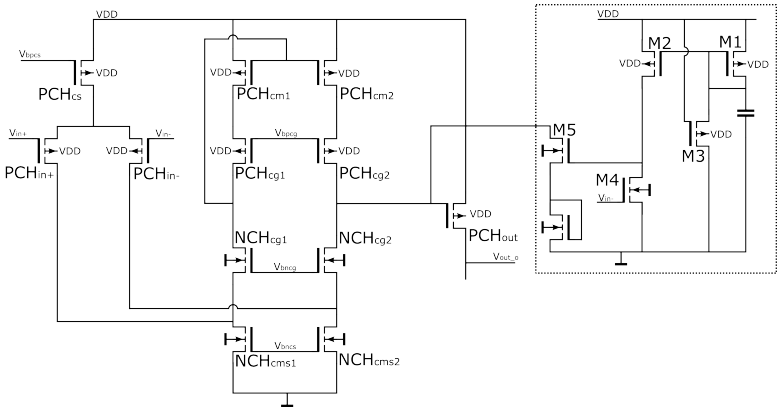

Figure 2.46: adding charge protector to start circuit of the output stage of core controller

2.3.11 Whole core controller

So far, the whole core controller's topology is decided and shown in Figure 2.47.

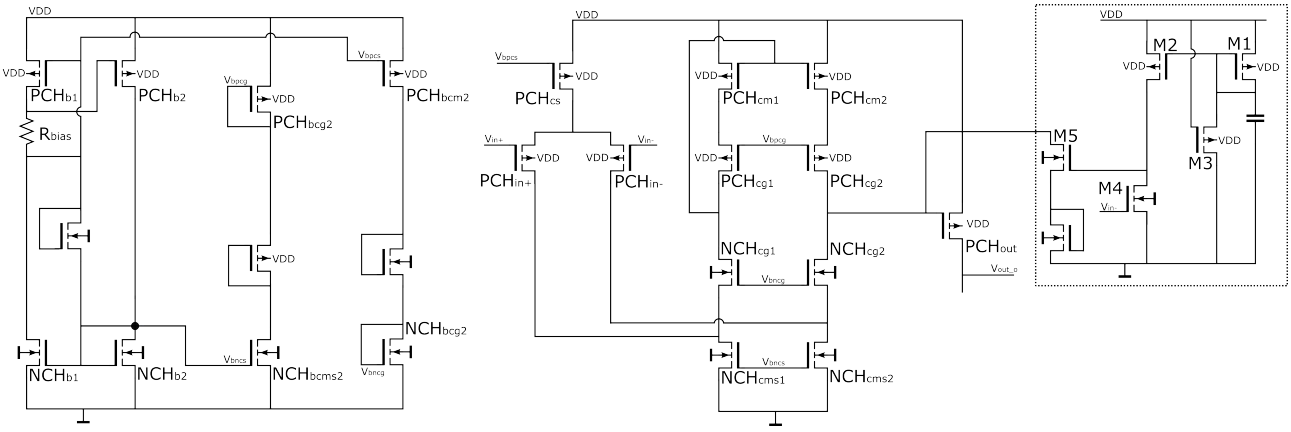

Figure 2.47: Whole core controller

2.3.12 Core controller: stability

An extra capacitor connected to the output stage is introduced into the circuit to ensure the core controller's stability, as shown in Figure 2.48. This is for making the dominant pole established by the cascode stage and the gate capacitor of the output stage more dominant. It also helps improve the circuit's power supply rejection(PSR) ability by establishing an AC coupling between the VDD and the gate of the output stage. It also establishes an RC low pass filter with the R_{out} of the cascode stage, which is important for the chopping design in future design steps.

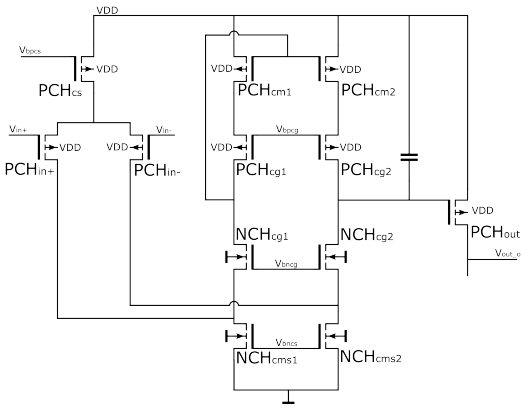

Figure 2.48: Adding an extra capacitor to ensure stability

2.3.13 Core controller: idea of chopping

The idea of chopping is for removing the input referred offset. Figure 2.49 shows how the chopping removes input referred offset in the time domain. The chopping will first move the input referred offset to the higher frequency. In the time domain, this occurs as square waves at the gate voltage of the output stage. No matter how much the input referred offset is, the square wave's average value is always the same. By sending this square wave to an RC low-pass filter, a triangle wave can be generated by making the capacitor not fully charged. Then with a sample and hold step, the midpoint of the triangle wave can be taken as the average value of the square wave.

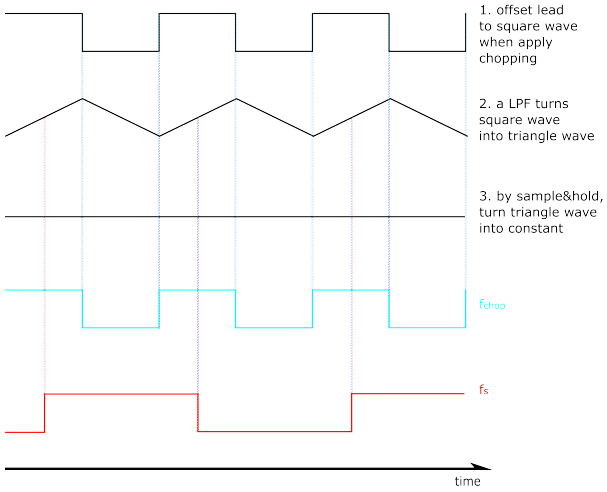

Figure 2.49: Chopping removes input referred offset in the time domain

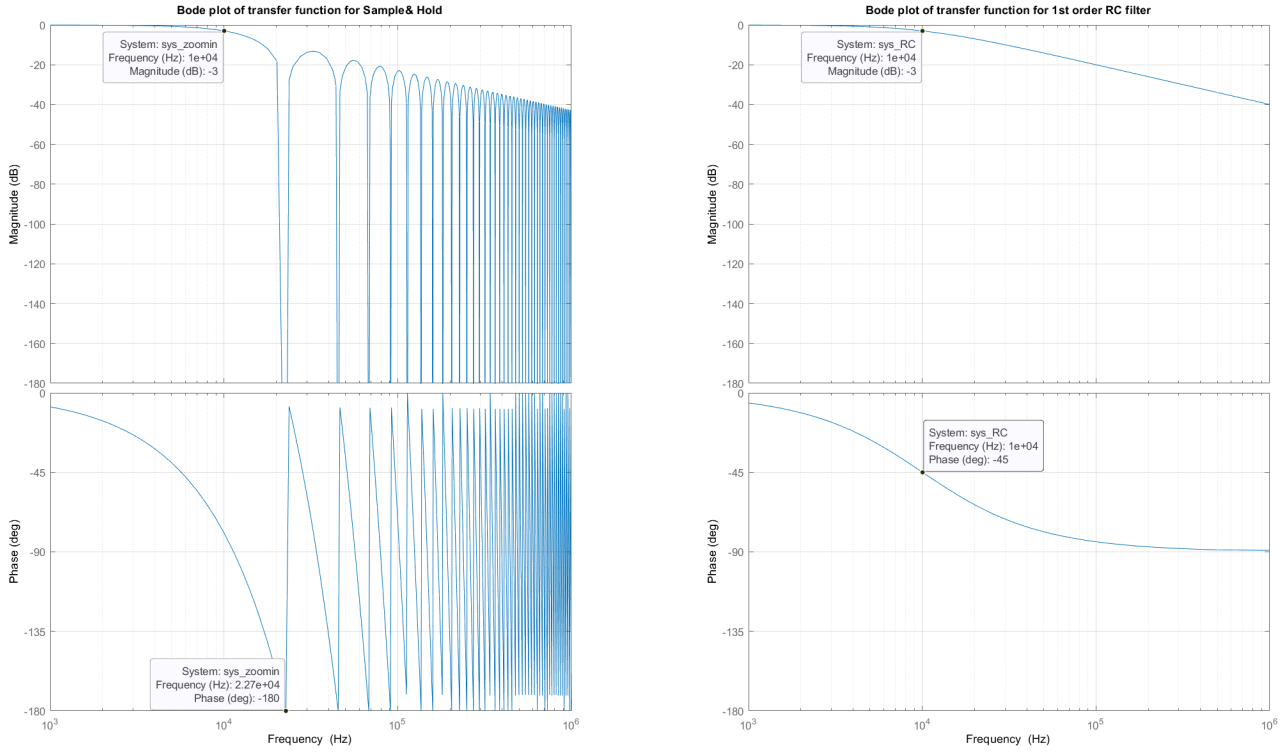

Figure 2.50: Bode plot of single pole low-pass filter and a notch filter that has the same $-3dB$ frequency at $10kHz$

From the view of the frequency domain, a low pass filter with an extremely low-frequency cut-off frequency can save the effort of the sample and hold step. But this would take too much area budget because a very big capacitor will be needed. Sample and hold can be used as a notch frequency domain filter. The magnitude of a single pole low-pass filter and a notch filter that has the same $-3dB$ frequency at $10kHz$ is shown in Figure 2.50. The bode plot shows that the notch filter can perform very good attenuation at the sampling frequency of the sample and hold step. Which is perfect for

removing the input referred offset at the exact frequency. And to realize such attenuation at the wanted frequency with only a single pole low-pass filter takes a large amount of area to establish an extremely low-frequency pole. So an acceptable-area RC low-pass filter combined with a sample and hold circuit would be a good solution for removing the input referred offset at the high frequency and keeping the signal at DC unchanged.

Besides that, chopping also helps with reducing the flicker noise. By chopping, the flicker noise can be moved to the chopping frequency and attenuated by the low-pass and notch filters by designing the chopping frequency to be 2 times of the sampling frequency. A detailed example and noise spectral density sketch can be seen in section 3.3.7. Note that this is only significant when the chopping frequency is much higher than the corner frequency of the noise. Otherwise, the chopper can only move the flicker noise lower than the chopping frequency to a higher frequency and can filter that out. In bandgap reference voltage design, noise canceling is not the key task for the chopping, so the design flow does not start with considering noise.

2.3.14 Core controller: chopping implementation

The first step is to implement the switches into the core controller and chop the input-referred offset to high frequency, as shown in Figure 2.51. The location of such choppers needs to be discussed, and the chopping nonidealities need to be considered. And the choppers need to cover the main input-referred offset contributors.

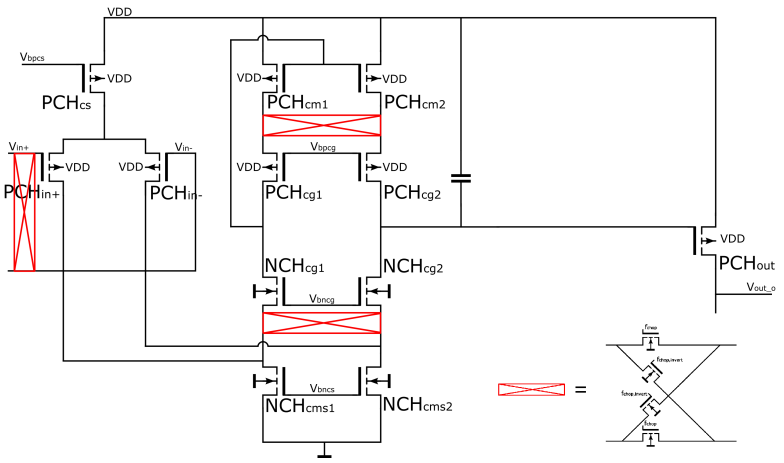

Figure 2.51: Implement choppers into the core controller

chopping nonideality: clock coupling and charge injection The clock signal for controlling the Finfets work as switches charge their gate capacitor. When the switches are open or closed, the charge leaks from the gate capacitor as extra current and flows into the input capacitor of the next stage. There are four solutions to this problem.

1. By applying two dummy switches in series with the original switch that has half width and the same length as the original switch which is shorted by connecting their drain and source and controlled by an inverted clock signal, as shown in Figure 2.52. When the original switch is open, the dummy switches on both sides will absorb the charges flowing out. And when the original switch is closed, they will supply the needed charge.

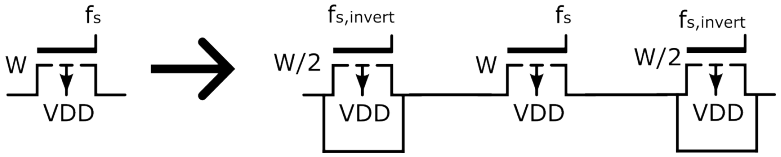

Figure 2.52: Applying dummy switches to the original switch

2. Another way of solving this problem is to increase the capacitance of the next stage. For example, by increasing the input capacitance of the input pair, the extra charge will not be able to charge the gate capacitor and generate a significantly high voltage. But this is also a trade-off with the input impedance, cause too much input capacitance leads to low input impedance(Z_{in}) at chopping frequency.

3. Designing the switches to have a less parasitic capacitor by minimizing the width could help(the length of the switches is preferred to be at the smallest possible value). Because less parasitic capacitance leads to less charge flows into the signal path. This is a trade-off with the switches' on-resistance(R_{on}), and too much R_{on} may require more Z_{in} to make it ignorable.

4. Lower chopping frequency could also help. The extra charges will flow into the signal path less frequently by lowering the chopping frequency, this also results in less extra voltage drop in the next stage. This is a trade-off with the capacitance used for forming the low-pass filter because the capacitor must not be fully charged. A lower chopping frequency requires more capacitance for a lower cut-off frequency in the low-pass filter and normally capacitors are quite area-consuming in modern techniques.

chopping nonideality: Input impedance As aforementioned, too much input capacitance may lead to too less input impedance at chopping frequency. The input impedance at chopping frequency will be as shown in (2.50). This problem can be solved by designing the chopping frequency to be low or designing the input capacitance to be less.

$$Z_{in} = \frac{1}{4 f_{ch} C_{in}} \tag{2.50}$$

chopping nonideality: Limited gain Even though the signal will be chopped back to DC, the unwanted offset will remain at f_{ch}. There are still some places where the signal runs at f_ch. Thus, the controller must perform high enough gain at f_{ch}. This can be promised by stronger inversion, which means more current is used for biasing. Lower chopping frequency could also help.

chopping nonideality: summary and design strategy The nonidealities and their trade-off is shown in table 4. Note that even though all the solutions have their advantage and disadvantage. Some disadvantages are so far away from the showstopper value or so less that they can be ignored. The area and power budget are also mentioned in the table.

From the tale, it can be told that most of the solutions are conflicts between different nonidealities. Only less chopping frequency benefits in all nonidealities. So it is always good to start by designing the chopping frequency to be as low as possible within an acceptable area. Because it only has the disadvantage of requiring more area. Thus start with finding a good f_{ch} with an acceptable extra capacitor area for the low-pass filter. Secondly, if the signal cannot enjoy enough gain at the f_{ch}, take more current budget to get enough gain at such frequency. Then design the switch width to be ignorable compared to the input impedance at f_{ch}. And apply dummy switches, because it has

Table 4: Chopping nonidealities and solutions to them($\surd$ = advanatge, $\times$ = disadvantage, $\circ$ = norelationship/ignorable).

nonideality	$f_{ch} \downarrow$	$C_{in} \uparrow$	$I_{budget} \uparrow$	Dummy switch	$W_{switch} \downarrow$
Clock coupling and charge injection	$\surd$	$\surd$	$\circ$	$\surd$	$\surd$
Input impedance	$\surd$	$\times$	$\circ$	$\circ$	$\times$
Limited gain	$\surd$	$\circ$	$\surd$	$\circ$	$\circ$
Circuit area	$\times$	$\times$	$\circ$	$\circ$	$\circ$
Power	$\circ$	$\circ$	$\times$	$\circ$	$\circ$

nothing to do with other nonidealities. Finally, if the input impedance is not big enough, adjust the input impedance or the switch area.

main input referred offset contributors The purpose of applying chopping is to remove the input-referred offset caused by mismatch variation in threshold voltage variation of Finfet. So it is necessary to consider which Finfets in the topology are the most contributors. Firstly, the main contributor is the input pair mismatch, this is normally the main area consumer of the controller design, but with chopping applied, it can be very small. Then the main contributors are the current mirrors. Unfortunately, to keep the polarity of the controller the same, only the mismatch from PCH_{in+} and PCH_{in-}, NCH_{cms1} and NCH_{cms2} are fully removed. There are still some remaining contributions from the PMOS current mirrors, which can be solved by designing the V_{gs} to be as big as possible and assigning an acceptable area to the current mirror. The common gate stages in the cascode stage are only for enhancing the R_{out}, so there is no need to worry about their contributions. An acceptable area of the common gate stages can already lead to a good result. Note that they are not the only mismatch contributors in the whole circuit. A section will discuss all the mismatch contributors and their contribution to the circuit in Section 3.3.

2.3.15 Core controller: notch filter

After having decisions on the choppers, the only thing left for the core controller is the notch filter. It can be realized by a sample and hold circuit works in ping-pong mode, as shown in Figure 2.53. The capacitors used in the circuit will influence the noise, but the low-pass filter connected to the V_{out} will dominate the noise. Note that the capacitor in the sample and hold circuit also needs to be much bigger than the input cap of the next stage. Otherwise, this may cause a voltage drop due to the sharing of charge.

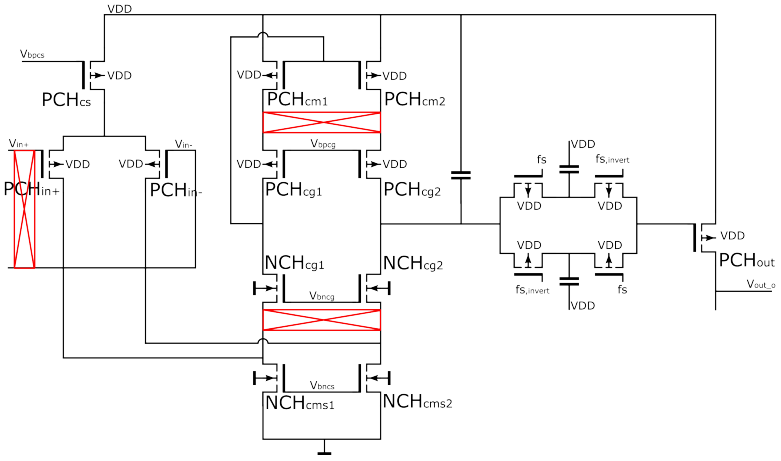

Figure 2.53: Implement notch filter into the core controller

2.3.16 Second controller: input stage

Different from the common mode voltage level of the core controller. Since the purpose of the second controller is to force the V_{out} on a resistor, which is a constant voltage of around $1.1V$, a pair of NMOS and an NMOS current source would be perfect for this case, as shown in Figure 2.54.

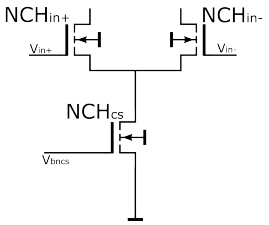

Figure 2.54: Input pair of the second controller

2.3.17 Second controller: output stage

The output stage of the second controller needs to be the same type as the output stage of the core controller because the purpose is to generate a current to bias a third BJT that has the same amount of current as one of the BJTs and keep the current being the same for all temperature. Thus the output stage will be connected to a temperature-independent resistor and a current mirror will be connected to the output stage, as shown in Figure 2.55. Note that PCH_{out1} and PCH_{out2} both requires big area, because the accuracy between their ratio will directly

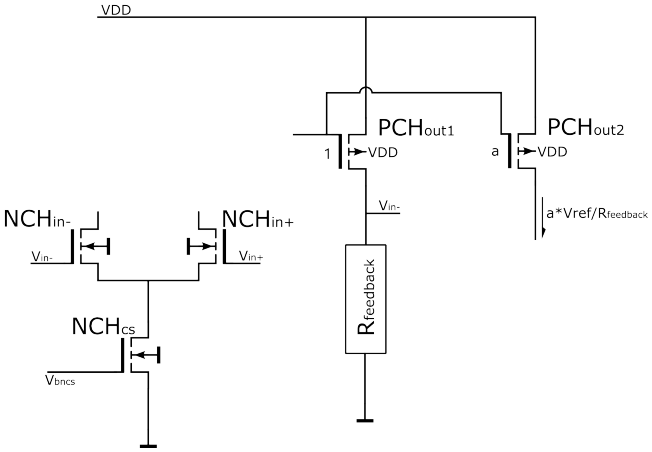

Figure 2.55: Output stage of the second controller

2.3.18 Second controller: current mirror

Now add the current mirror to take good advantage of both input pair and polarity, as shown in Figure 2.56. The area deigning of the Finfet is in the same way as the Core controller.

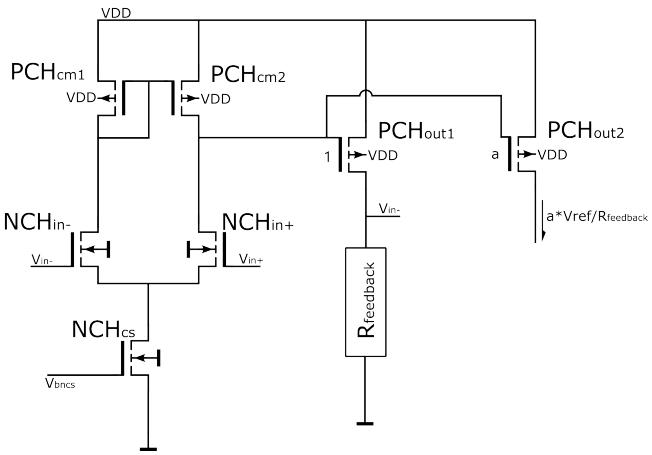

Figure 2.56: Current mirror of the second controller

2.3.19 Second controller: start circuit

The input pair of the second controller is connected to the V_{out}. Thus only when the V_{out} is high enough, the output stage of the second controller can be biased well. This means that the second controller will only start settling when the core controller is midway through its settling. Thus the V_{out} will have two steps in settling behavior, as shown in Figure 2.57. Because the V_{out} will only locate at its final value when both controllers are at their steady states.

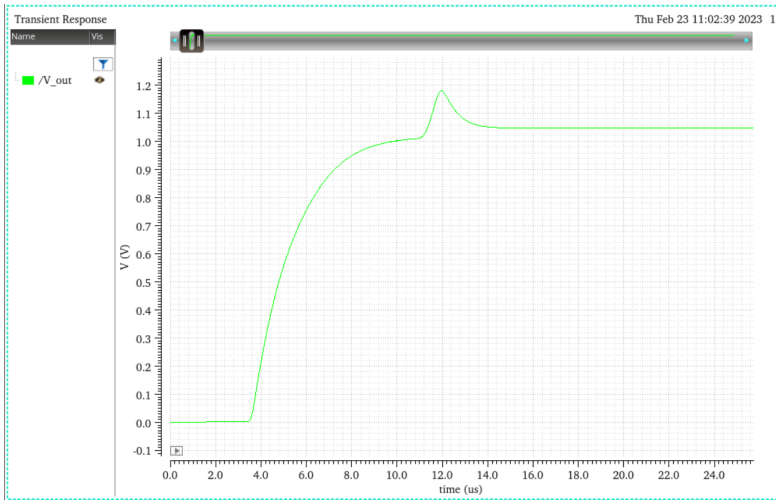

Figure 2.57: Two-step settling behavior

A start circuit is applied to the second controller to avoid this result. Note that this start circuit is not necessarily needed in the circuit because the two-step settling behavior is not a stability problem, and there is no such 0 current occasion to avoid. As shown in Figure 2.58, it's the same topology used in the core controller, and this time the start circuit is started when the V_{out} is not high enough, the gate capacitor of the output stage will be precharged until the V_{out} rises to an acceptable level, then the feedback will take care of the gate voltage. This saves time for the first stage to charge the output stage, which should help avoid the two-step settling behavior. The final settling behavior will be shown in Section 3.

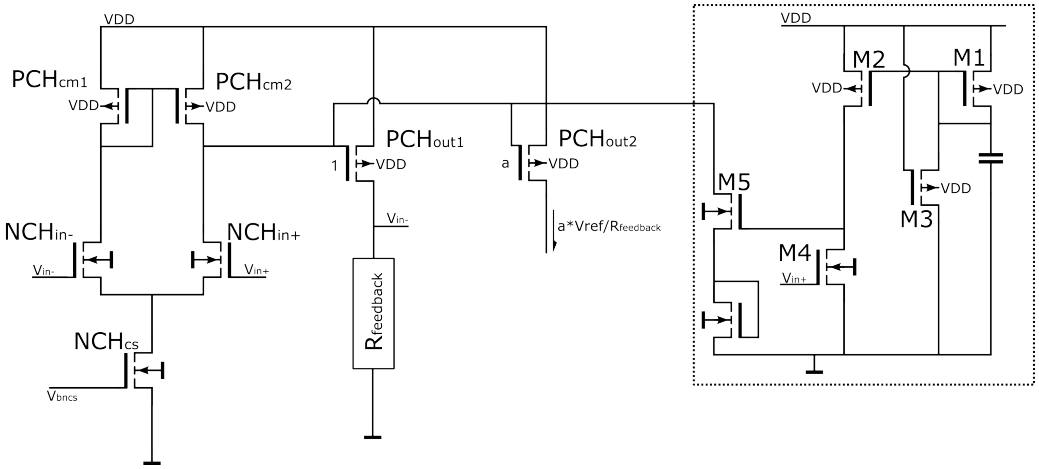

Figure 2.58: Start circuit of the second controller

2.3.20 Trimming resistor

The only component left now is the trimming resistors, as shown in Figure 2.59. The accuracy of each step is not highly required in such a trimming resistor. No extra area is needed for the accuracy. The minimum area for each resistor should be good enough. Thus a binary structure would be good enough for such design.

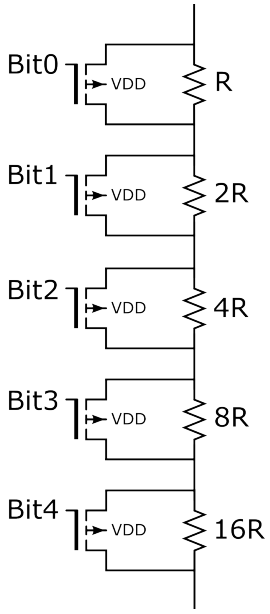

Figure 2.59: Trimming resistor

The first step is to find out the value for each step of resistors. This can be told from the worst case of Monte Carlo simulation, for example, the one in Figure 3.5. Note that it will be better to do this after finishing all the other designs including the controllers. Because even though the offset from the controller does not perform in a PTAT way, they can still be partly removed, as shown in Figure 2.60. For example, the offset is ΔV, when the V_{out} is perfectly trimmed to the ideal value at the middle temperature($55°C$), then the residual error should be $(\frac{150+273.15}{55+273.15} - 1)\Delta V \approx 30\%\Delta V$. Note that the designer should not count on using PTAT trimming to remove the offset, because it's also not perfectly constant in practice.

The second step is to decide to use which type of Finfet to realize the switches. In this case, the trimming resistor will have roughly $1V$ on both sides. And for a better-performing triode region Finfet purpose, it's better to use more V_{gs}. Thus it's better to use a PMOS, this could promise roughly $V_{gs} = 1V$.

Then with the designed resistor value, find out how much Length is required for each Finfet. Making R_{on} at least smaller than 10% of the resistor it needs to short is preferred.

2.3.21 Process & Mismacth variation

The design steps so far are only for realizing the behavior of the circuit. Some components may contribute to the variation behavior of the whole circuit. Thus the designer shall run multiple times of Monte Carlo tests with only considering the mismatch of each component. This could help with checking the contribution of each component and planning area for them. Note that since the circuit behavior is already promised. In this step, the designer shall only consider increasing the transistors or resistors' width and length by the same ratio.

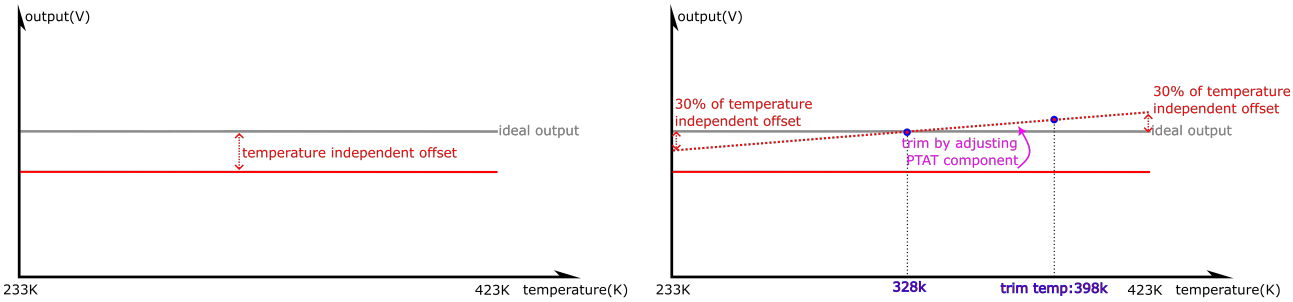

Figure 2.60: The offset is partly trimmable

2.3.22 Conclusion for schematic level

The aforementioned are the steps for designing the schematic level. The detail around design parameters and the circuit result will be shown in Section 3.3. The design flow in this section shall lead to a minimum area and an acceptable power budget for such a bandgap reference voltage source, which is wanted in this project.

2.4 layout level

The components that consume most of the area will be listed. The reason why they require so much area as well. The points around such components must be kept an eye on while designing the layout will also be discussed. The order is related to the area captured by the components, from high to low.

2.4.1 Capacitor for the low-pass filter in the core controller

This is the biggest capacitor and also the biggest component used in the whole circuit. Due to the requirement for an extremely low-frequency cut-off frequency, this capacitor will be very big. Such cut-off frequency can also be realized by increasing the R_{out} of the cascode stage, but the capacitor will still occupy a significantly large area. And the design of this capacitor should be done after finishing designing the core controller. The area of such a capacitor is decided by the maximum acceptable capacitor area in the project.

2.4.2 Capacitor for the low-pass filter connected to the output of core controller

This is the second biggest capacitor in the circuit. Due to the noise requirement, the capacitor has to be very big. The noise behavior directly decides its area.

2.4.3 Current mirror in the second controller

This current mirror is the only current mirror included in the signal path. Thus its accuracy must be high enough, which needs a large area if the designer does not apply extra techniques to increase the accuracy. Fortunately, in the structure mentioned in this report, this is the only current mirror used in the signal path. Thus avoiding using current mirrors in the signal path could save a lot of area. The mismatch between the current mirror is the point that needs to be considered while designing the layout.

2.4.4 BJT

The BJT will contribute most of the thermal noise in the circuit. The spectral density will be $\frac{2(kT)^2}{qI_c}$. This means that the current planned to use to bias the BJTs will directly dominate the noise result. But due to the design decisions on the bias density for the BJTs, the BJTs' area needs to increase by the same ratio if the designer wants to increase I_c. Thus the area of BJT will also be significant. Increasing the capacitor area in the low-pass filter connected to the output of the core controller can also help. So it's a trade-off between the area of the BJT and the capacitor area. The mismatch between the BJTs is the point that needs to be considered while designing the layout.

2.4.5 Core controller(except the capacitor)

Normally, the input pair is the most area-consuming component in a controller design. This is because it's the first stage in the controller, which would dominate the noise and the input-referred offset. As mentioned in section 2.3.7, only the flicker noise is dominated by the input pair's $width \cdot length$ area. And chopping techniques will attenuate such low-frequency noise, the input-referred offset contributed by the input pair in a more area-effective way. For example, an input pair increases by 25 in area can decrease the V_{th} mismatch by 5, but the same area of a capacitor may easily reach a reduction by 20 times in input-referred offset. Thus with the chopping technique, the area occupied by the core controller(except the capacitor) decreases dramatically. The mismatch between the input pair and the current mirrors is the point that needs to be considered while designing the layout.

2.4.6 Bias circuit

Due to the settling behavior among such a 5σ yield. The bias circuit needs to be big in case of failing to start. The mismatch between the current mirrors is the point that needs to be considered while designing the layout.

2.4.7 Trimming resistor

Due to the smallest step of the resistor, the transistors used in the trimming resistor must have a big enough width to promise a low R_{on}.

2.4.8 Second controller(except the current mirror)

Due to the purpose of the second controller, its input-referred offset will not influence the V_{out} too much. So the area except the current mirror will not be too much. The mismatch between the input pair and the current mirrors is the point that needs to be considered while designing the layout.

2.4.9 Conclusion for layout level

Normally, designers would think that the circuits with chopping techniques would take more area because designers take the extra capacitor as extra stuff to the circuit. But actually, it is not true. The designer should take away the area spent on the transistors for less mismatch and flicker noise, then spend that on the extra capacitor. This means that chopping is not always a more area-costing strategy. The floor plan of this project can be an example to prove this. The detail of the area occupied by each component will be in Section 3.3.

3 Design decision

Based on the analysis in chapter2, design decisions will be made based on the detailed occasions. This chapter will follow exactly the same order as chapter2 because decisions at each step are directly made after finishing the analysis.

3.1 Principle model level

As mentioned in chapter 2.1.3, the first error to consider is the errors from BJT. The errors directly connect to BJT devices' physical quantities are saturation current variation, current gain, base resistance, and curvature error.

The saturation current variation can be ideally calibrated. The curvature error can also be calibrated. But the current gain and base resistance will always influence the $V-T$ behavior of the BJT by adding nonidealities to the $V-T$ of diode-connected BJT and making the relationship non-first-order.

This means that the first break-point is the $\beta-T$ behavior among different bias currents (current gain) and the BJT's $V-I$(base resistance) behavior in the given process. And current gain leads to the least bias current density, and base resistance leads to the most bias current density.

Two different types of processes can be chosen, C040GF and 16FFC, which are 40nm and 16nm. The 16FFC is more advanced and will be used in the integrated circuit in the coming products in NXP. Thus if 16FFC is sufficient, it would be better to continue with designing in 16FFC.

3.1.1 $\beta-T$ behavior

According to the theoretical analysis from chapter 2.1.5 and the result in [2]. It can be concluded that with $\beta > 5$ for all temperatures in the target range, the influence of the current gain will be negligible.

Two different types of BJT are available in 16FFC, NPN, and PNP. For every kind of BJT, two different versions of Emitter area can be chosen: Emitter area1$= 0.682\mu m \cdot 4.166\mu m$ and Emitter area2$= 1.364\mu m \cdot 4.166\mu m$. In the test bench, the BJTs are connected in Figure 2.7 and biased by different amounts of current. The $\beta-T$ relation of NPN and PNP in 16FFC under different bias current densities are plotted in Figure 3.1 and Figure 3.2.

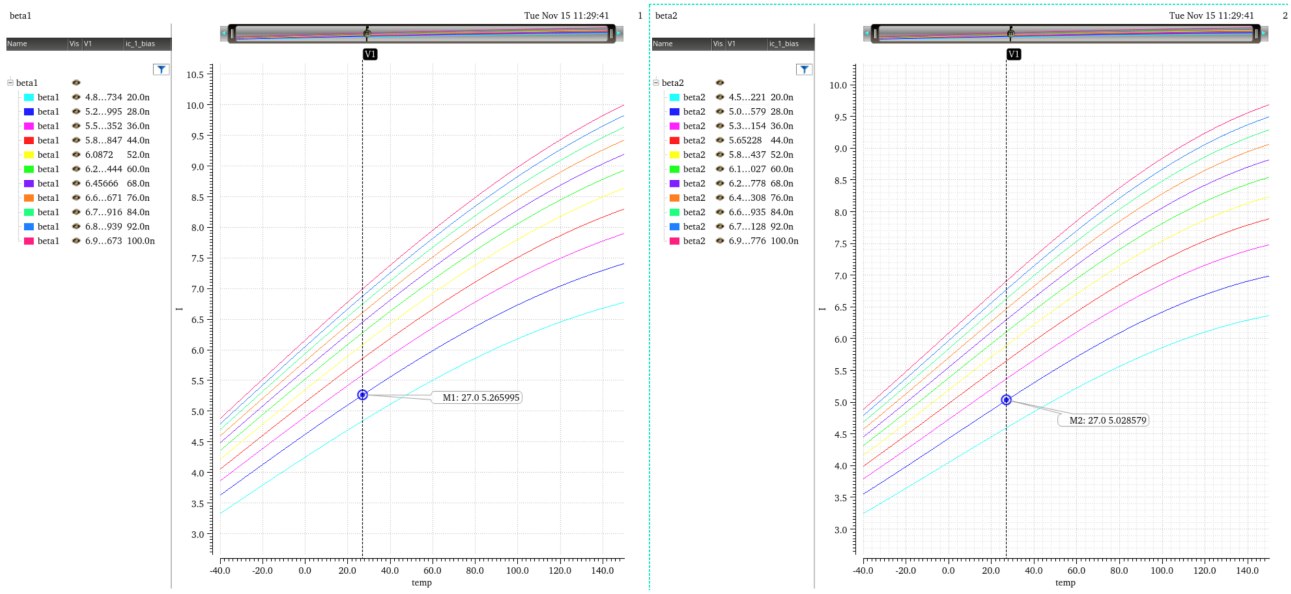

Figure 3.1: $\beta - T$ behavior of NPN BJT in 16FFC under different bias currents. beta1 represent Emitter area1, beta2 represent Emitter area2

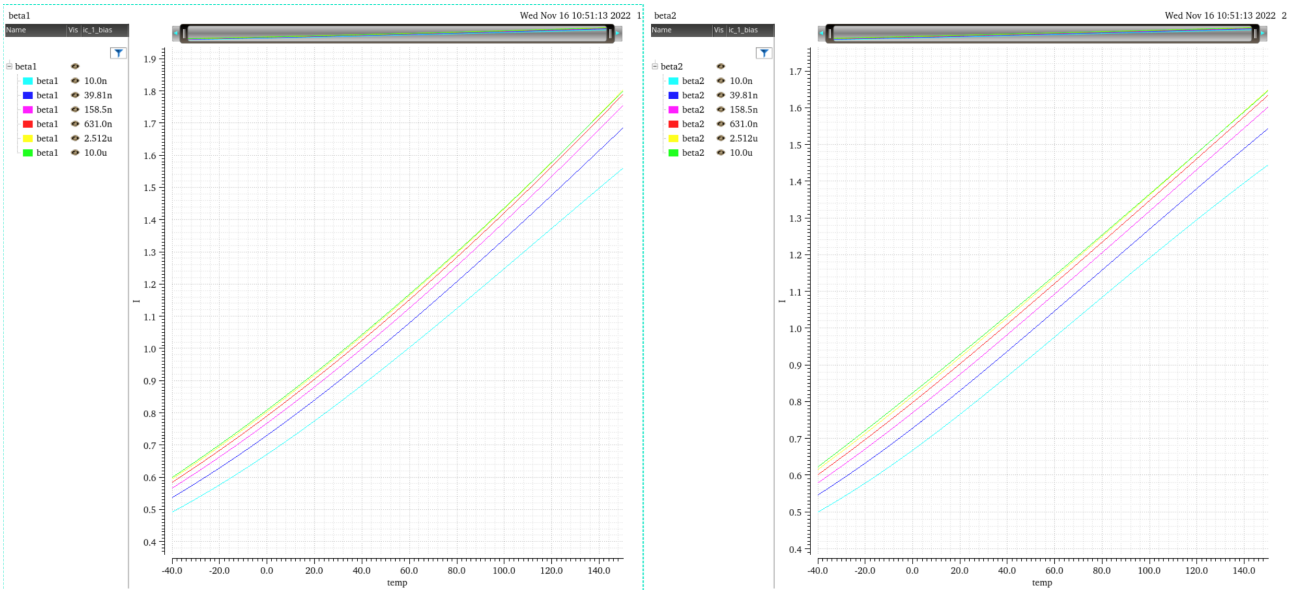

Figure 3.2: $\beta - T$ behavior of PNP BJT in 16FFC under different bias currents. beta1 represent Emitter area1, beta2 represent Emitter area2

From the $\beta - T$ behavior, it can be told that no matter how much current is injected into PNP BJT in 16FFC, the current will not be sufficient. And for NPN, a sufficient current gain can be achieved easily with a small current. This means that NPN BJTs in 16FFC are sufficient in current gain as long as more than $30nA$ of current is injected into every Emitter area$= 0.682\mu m \cdot 4.166\mu m$. This also defines the least value of bias current density for BJTs.

3.1.2 $V-I$ behavior

Then check the $V-I$ behavior of the NPNs. Since the $V-I$ behavior of a BJT can be defined by V_G at $0k$ and V_{BE} at T_r, discussing the behavior at T_r would be already enough, in this case, $T_r = 27°C$. The purpose is to avoid the Base resistance influencing the linearity of the $V-I$ behavior. Thus checking the saturation current varying with the bias current can perfectly tell the linear region of $V-I$ behavior, as shown in 3.3.

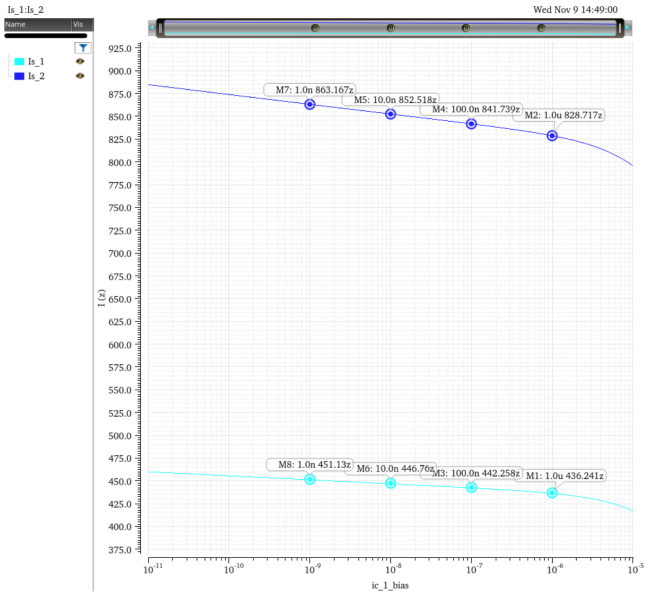

Figure 3.3: saturation current-bias current density behavior of NPN BJT in 16FFC at $27°C$

The Figure shows that when the bias current for each Emitter area$= 0.682\mu m \cdot 4.166\mu m$ is more than $1\mu A$, then the Base resistance will make the saturation current no longer linear. Since this conclusion is based on the bias current density and Base resistance decrease when the Emitter area increases. Therefore, it can be concluded that the bias current density shall never be more than $1\mu A$ for every Emitter area$= 0.682\mu m \cdot 4.166\mu m$ at $27°C$.

3.1.3 Conclusion of concept principle model level

At the principle model level, the most important thing is to name an acceptable range for the CTAT voltage generator(BJTs), which promises the devices to work with good linearity. And the acceptable range is for every Emitter area$= 0.682\mu m \cdot 4.166\mu m$ of NPN in 16FFC at $27°C$, always inject current in a range of $30nA$ to $1\mu A$. this also points out that 16FFC is sufficient for design bandgap voltage reference. According to the conclusion from section 2.2.3. The max and min boundaries can be chosen as the design values.

3.2 Concept level

3.2.1 Design decisions

In the last section, the bias current density range is chosen ($30nA$ to $1\mu A$). According to the noise analysis in section 2.1.8. $\frac{I_{s1}}{I_{s2}} = \frac{8}{1}$ and $\frac{I_{c1}}{I_{c2}} = \frac{1}{4}$ is designed. The noise behavior of this setup should be better than $I_{c1} = I_{c2}$. $\frac{I_{s1}}{I_{s2}} = \frac{8}{1}$ can be realized by a square-located layout, which would lead to less

mismatch. As shown in Figure 3.4, BJT_2 can represent the average value for the variation among two dimensions, so BJT_1 matches BJT_2 better. This also works for conditions like$1:24, 1;48, 1:80, ...$, but considering the acceptable current density range is $\frac{30nA}{1\mu A} = \frac{1}{33.33}$, $1:8$ in area and $1:4$ in current is chosen. Thus $\frac{I_{c1}}{I_{s1}} \cdot \frac{I_{s2}}{I_{c2}} = \frac{1}{32}$.

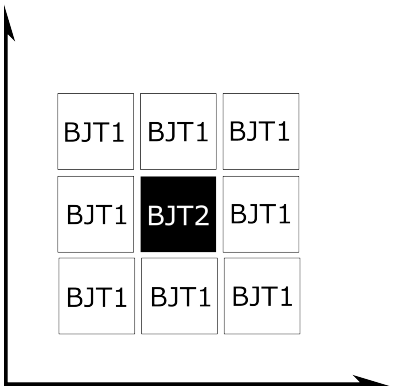

Figure 3.4: Square-located layout for BJTs

The ratio between R_1 and R_2 can be calculated, which is $\frac{R_2}{R_1} = \frac{I_{c1}}{I_{c2}} = \frac{1}{4}$. The absolute value of R_3 can be defined by letting $\frac{V_t \cdot ln(32)}{R_3} = 30nA$ at $27°C$.

According to the knowledge from section 2.1.7, the ratio between R_1 and R_{c1} and ratio between R_2 and R_{c2} needs to be slightly more than $\eta = 4.27$, so take 5. Because if the coefficient is too much, the mismatch from the second opamp will be amplified more times. Since the constant biased BJT needs to have the same bias current density at $27°C$ and I_c and I_s shall be multiple times more than NPN2. Take four times in design. Thus the resistance defines the constant current, and the I_s of the constant biased BJT can be worked out.

3.2.2 Circuit behavior

Four 5σ Monte Carlo simulation results with different devices replaced with ideal devices are shown in Figure 3.5. The red curves represent the max cases, the blue curves represent the mean value, and the yellow curves represent the min cases. The figures in the first line are the absolute value of V_{out}, and the figures in the second line are the values of the error of the curves in the first line.

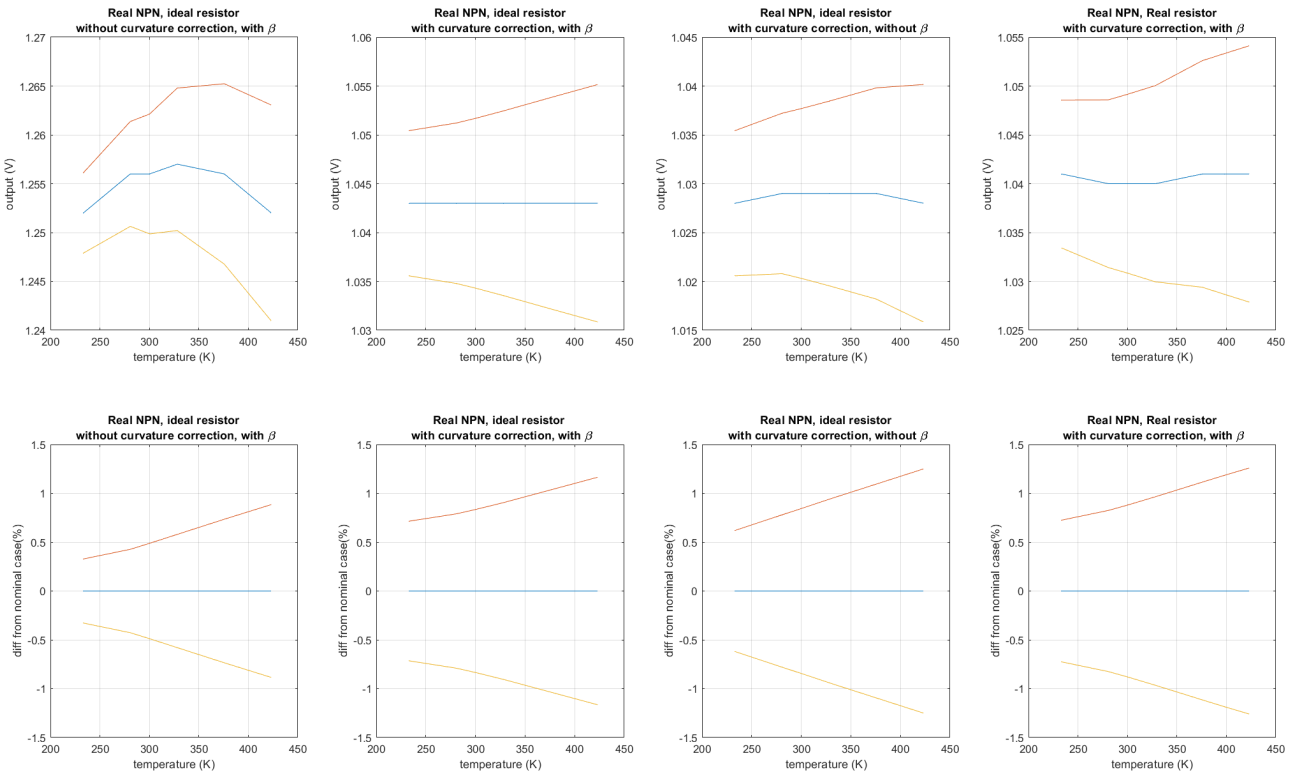

Figure 3.5: Circuit behavior of Max, Min, typical case

Real BJT, ideal resistor, without curvature correction, with current gain The first case is for showing the influence of the curvature. Since the curvature error is a physical behavior, if there is no such curvature correction for it, then there will be around $5mV$ curvature error in the V_{out}. And curvature error is roughly the same for all cases. This is why the curvature error cannot be observed when plotting the error between max, min, and mean.

Real BJT, ideal resistor, with curvature correction, with current gain The second case shows the result when the curvature correction part is added to the circuit. It can be seen that with curvature correction, the variation of curvature error is within $1mV$

Real BJT, ideal resistor, with curvature correction, without current gain The third case shows the result when no current gain influences the I_c of BJT. The error plot shows three straight curves, and the previous two cases do not; without current gain influencing the I_c, the second-order error in V_{BE} is removed. By comparing the second and third cases, it can be concluded that the second-order error brought by the current gain is ignorable.

Real BJT, real resistor, with curvature correction, with current gain The Fourth case is realized by replacing the resistors in the second case with the real ones. It can be seen that the variation is slightly more, but compared to the variation caused by the BJTs, they are very small.

Prediction of the possibility of trimming The trimming is realized by adjusting the coefficient of the PTAT part. In practice, this is done by one temperature point trimming at $125°C$. This

Table 5: Result before and after the ideal trimming

temp	mean	std	%	$+5\sigma$	ideal result after trimming at $125°C$
$-40°C$	$1.041V$	$1.514mV$	0.145%	$7.57mV$	$0.35mV$
$7.5°C$	$1.040V$	$1.717mV$	0.165%	$8.585mV$	$-0.106mV$
$27°C$	$1.040V$	$1.832mV$	0.176%	$9.16mV$	$-0.135mV$
$55°C$	$1.040V$	$2.009mV$	0.193%	$10.045mV$	$-0.117mV$
$102.5°C$	$1.041V$	$2.32mV$	0.223%	$11.6mV$	$-0.033mV$
$125°C$	$1.041V$	$2.466mV$	0.237%	$12.33mV$	$0mV$
$150°C$	$1.041V$	$2.625mV$	0.252%	$13.125mV$	$0.02mV$

means knowing how much voltage headroom needs to be trimmed at $125°C$, the magnitude of voltage changes in V_{out} at other temperatures can also be calculated, as shown in (3.1). By doing this, the residual error from BJT and resistors after ideal trimming can be predicted. Since they are all temperature-dependent errors, the residual error should be small enough to be ignored. The result before and after the ideal trimming is shown in table 5.

$$V_{trim,T} = \frac{T + 273.15}{125 + 273.15} \cdot V_{trim,125°C} \tag{3.1}$$

3.2.3 Conclusion for concept circuit level

The errors being analyzed are mainly the errors from the BJT and the resistors. The error before and after one temperature trimming shows that it is true that the discussed errors do follow the PTAT behavior. This means that in the given 16FFC process. The BJTs are able to be used for building a bandgap reference circuit. And with a single temperature trimming. The variation from the BJT and resistors shall be able to be removed.

3.3 Schematic level

In this section, the detailed design decisions on the schematic level and the result will be shown. The design decision follows the design order from section 2.3. Start with the modeling of the transistors. Then the design for the low-pass filter for the triangle wave. And in the final end with the behavior of the circuit. The result of the behavior should answer the requirement lest in table 1.

3.3.1 Modeling of PMOS and NMOS

As mentioned in the previous chapter. With an acceptable current budget and a good gm/ID model for the transistors. The designer could make decisions on the transistor area. In the projects, the transistors which contribute their gm are designed to have effective gm/ID. And the transistors work as current sources are designed to as less W/L as possible. Thus their gm will be as small as possible. Figure 3.6 shows the gm and noise behavior of PMOS and NMOS in the given process. The model is built under $I_{DS} = 5\mu A, L = 240nm, Fin = 20$ and with varying only the number of fingers. Interestingly, in 16FFC, the gm/ID model of PMOS and NMOS are similar. This means that for the same quiescent current, the designer does not need to spend $2 \sim 3$ more gate area for PMOS in 16FFC.

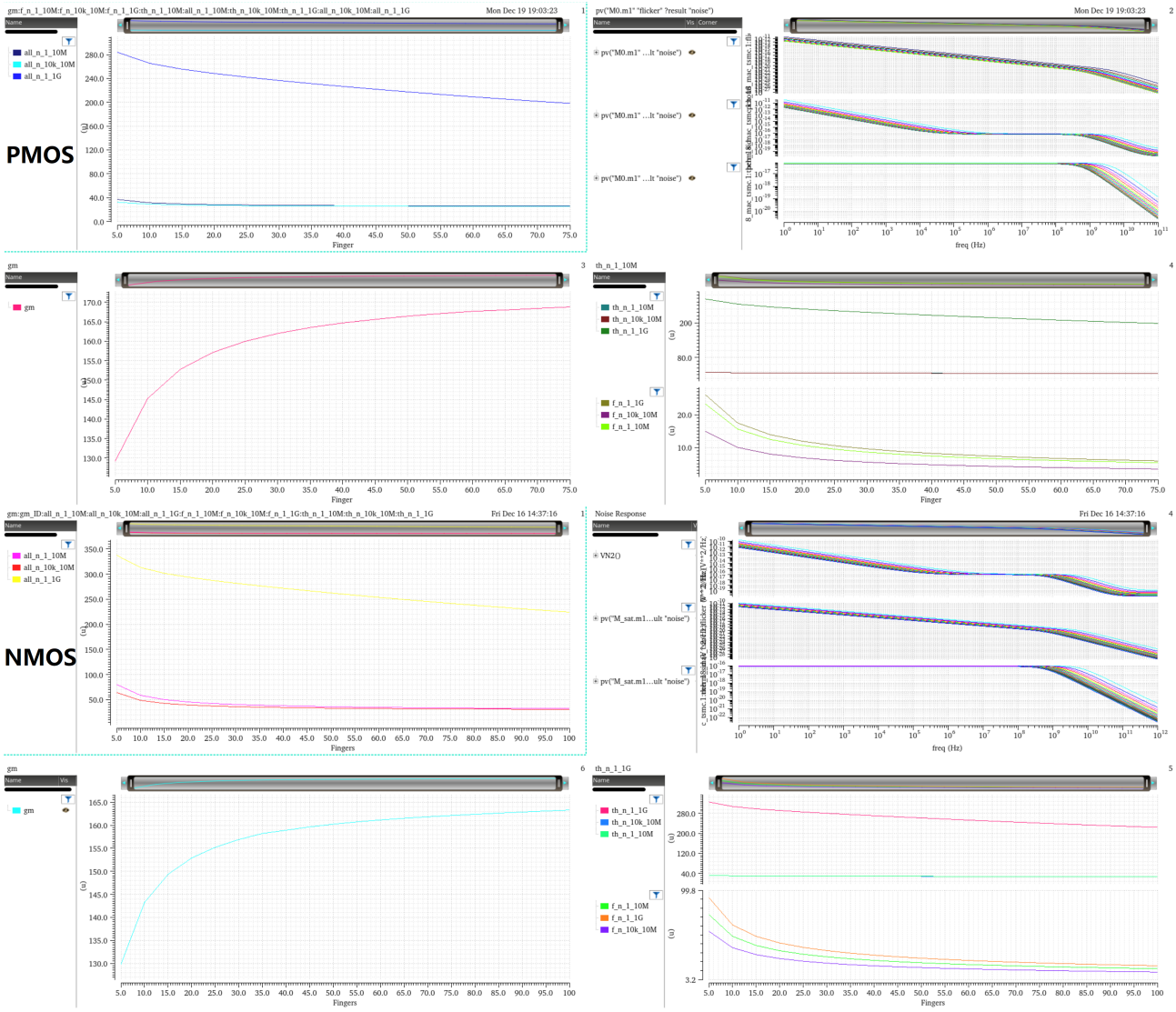

Figure 3.6: Transistor modeling

The Figure shows that for every $5\mu A$ quiescent current, a transistor with $L = 240nm, Fin = 20, Finger = 30$ leads to a good gm and noise result. Then it's time to move to the next level, assigning the current budget for each branch in the controller.

3.3.2 Current budget

The core controller is the most power-consuming component. The current budget of it is shown in Figure 3.7. The designer can have a good starting point for the transistors' area from the current budget and the conclusion from the last section.

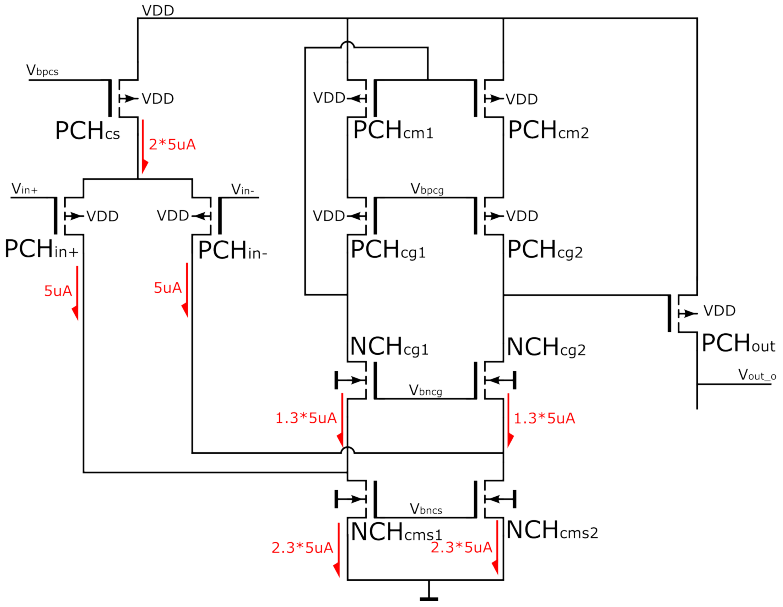

Figure 3.7: Current budget of the core controller

For example, for common gate stages in the second branch. The starting point is $Finger = 1.3 \cdot 30 = 39$ because the designed quiescent bias current is $1.3 \cdot 5\mu A$. And the reason for 1.3 is that when one of the input pairs is closed, there could still be some current left for the cascode branch.

3.3.3 Design for the low-pass filter in core controller

From the experience of previous designs. A $50pF$ capacitor would be already very big for such a bandgap reference circuit. So the design started with a $50pF$ extra capacitor attached to the circuit, as shown in Figure 3.8. The loop gain without and with the extra capacitor is shown in Figure 3.9, The deep blue one is the one with the extra capacitor.

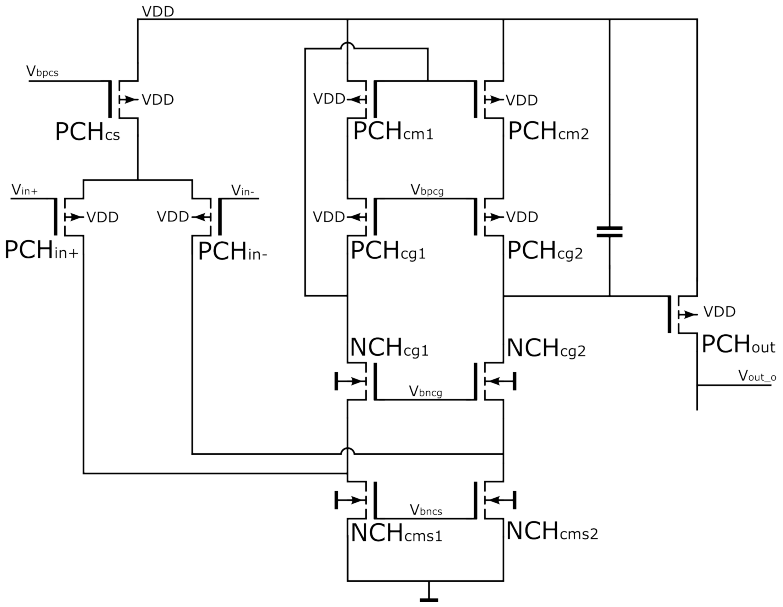

Figure 3.8: Adding an extra capacitor to ensure stability

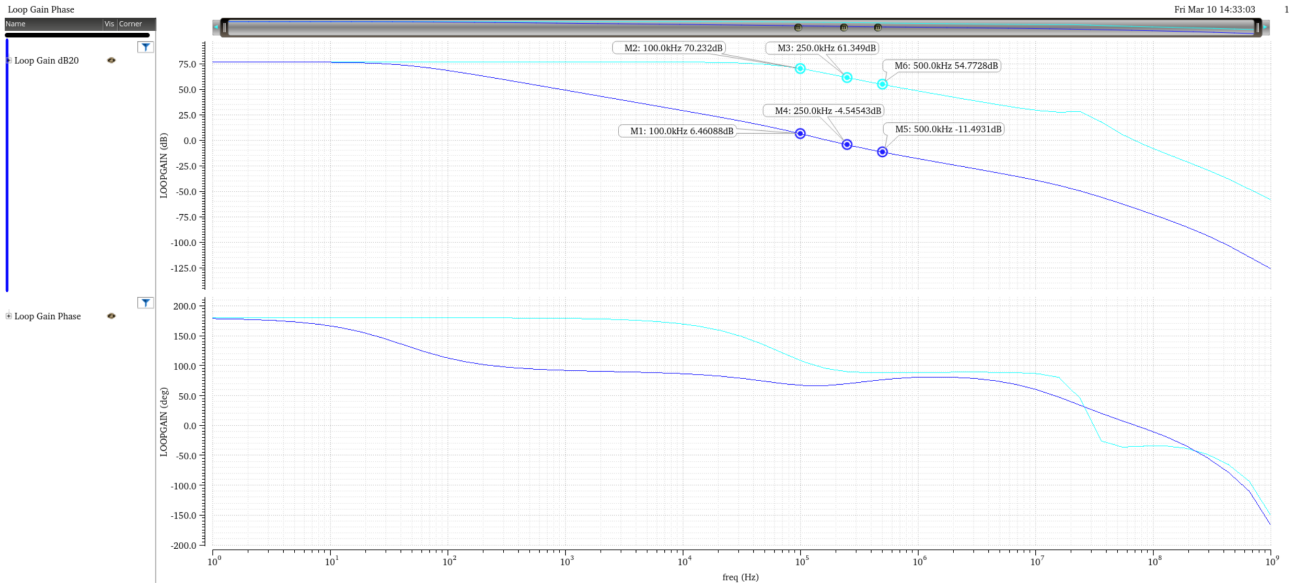

Figure 3.9: Loop gain without and with the extra capacitor

The point of realizing a low-pass filter is to let the unwanted offset be chopped to a high frequency and not let it fully charge the capacitor in the low-pass filter, which leads to the wanted triangle wave. For such purpose, the chopping frequency must be chosen to have a loop gain lower than $0dB$. In Figure 3.9, $250Hz$ meets the requirement. The design started with $C_{extra} = 40pF$ and $f_{ch} = 300kHz$ for safety headroom.

3.3.4 Chopper area

The first step is to have a rough model for the R_{on}, as shown in Figure 3.10. This is PMOS and NMOS with $L = 135nm, Fin = 20, Finger = 120$ and $|V_{GS}| = 1V$, $V_{DS} = 0.1V$. With such a model, the designer can directly get to the area leading to the wanted R_{on}.

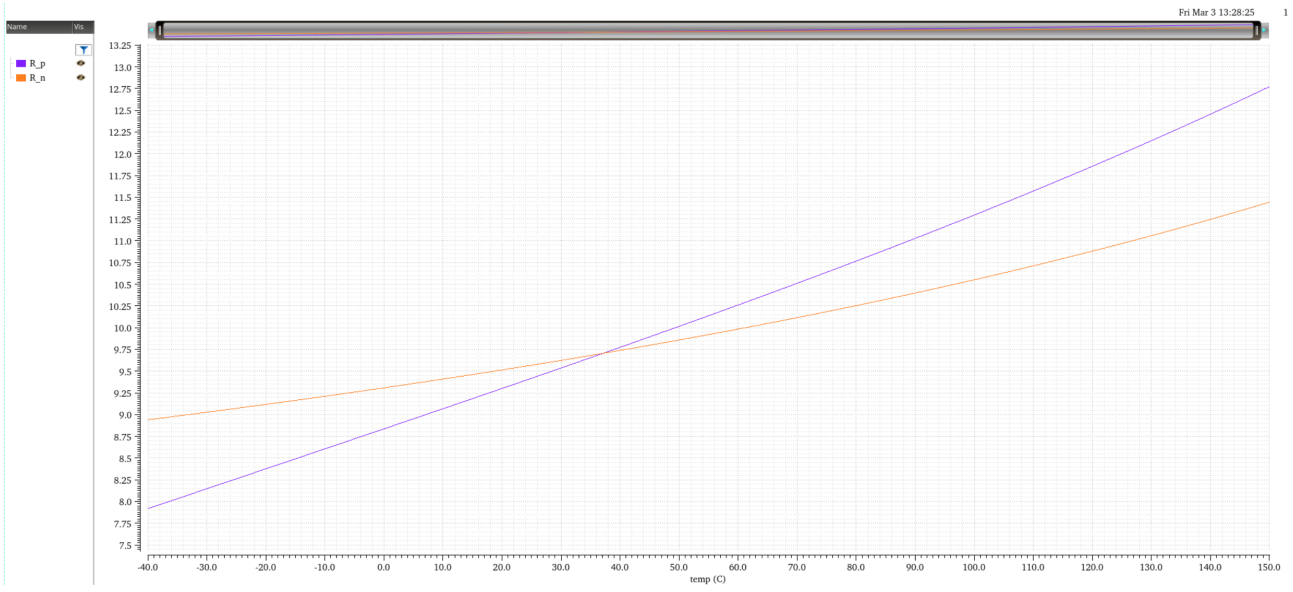

Figure 3.10: R_{on} of PMOS and NMOS at different temperature

With knowing the controller and the chopping frequency. By applying the gate capacitance of the input pair and the chopping frequency to (2.50). In this project, $Z_{in} = \frac{1}{4 \cdot 300kHz \cdot 100fF} = 8.3M\Omega$. 1% of this value is $83k\Omega$, which is quite easy to achieve. So in this project, the chopper is designed to have $Finger = 4, Fin = 20, L = 135nm$. This leads to roughly $R_{on} = 400\Omega$ and a very small area.

3.3.5 Trimming resistors

After finishing the chopper design, it's time to design the trimming resistor and the switch for the trimming bits. By checking the Monte Carlo results and trying to vary the resistor $R_1||R_2$ in Figure 3.11 to make the 5σ worst cases within the acceptable range. The wanted resistance range for the trimming resistor is found. In this project, the process variation at $150°C$(max process variation temperature) in 5σ is $\pm 10mV$. $10mV/2^5 = 0.3125mV$ means that 5-bit trimming is good enough for this project. Thus dividing the found resistance range by 32 leads to the wanted resistance; this project's value is 150Ω. To make sure the switch will not influence the resistance, target for $R_{on} = 15\Omega$.This leads to Finger=120.

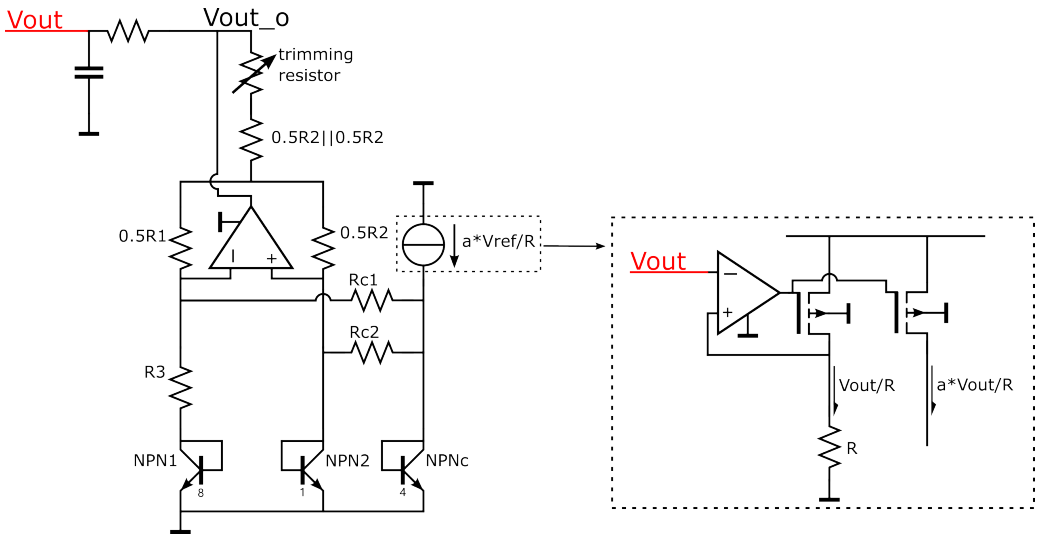

Figure 3.11: Replacing the nullors with controllers and add low pass filter

3.3.6 Startup(settling) behavior

In practice. The supply voltage will be ready first, and then the reference voltage circuit will be enabled after some time. This enabling activity will be executed within $1ns$. Thus the startup behavior is checked in different processes, VDD, and temperature corners. As shown in Figure 3.12. The VDD is enabled after $20\mu s$. There are still some corners that have two steps in the startup. This is because the bias circuit takes longer at some corners than the start-up circuit to settle(e.g. ss corner at $-40°C$). The first step is caused by the precharge of the startup circuit. The result shows that the circuit can start in $160\mu s$ in all conditions.

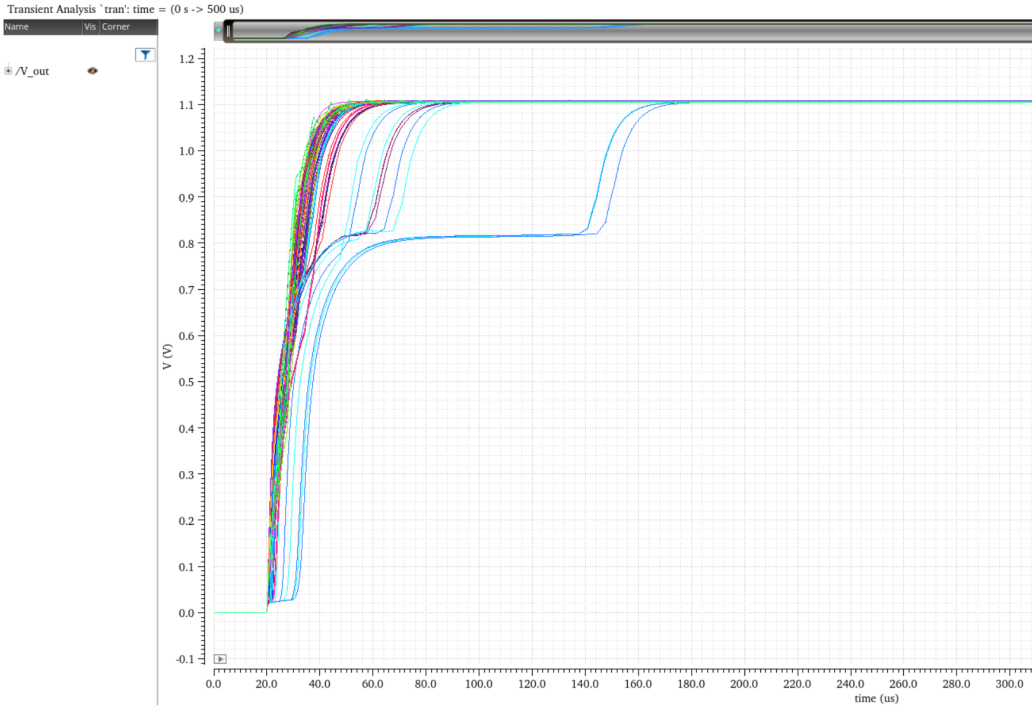

Figure 3.12: Startup(settling) behavior in all PVT corners

3.3.7 Noise behavior with chopping

As mentioned in the literature review, the flicker noise shall be able to be moved to a higher frequency(f_{ch}). And in this project, an example can be shown at 27°C and at the typical corner. Figure 3.13 shows the noise spectral density without chopping. Figure 3.14 shows the noise spectral density with and without the notch filter. Note that a low-pass filter is attached to the output, so the figure has no flat noise floor.

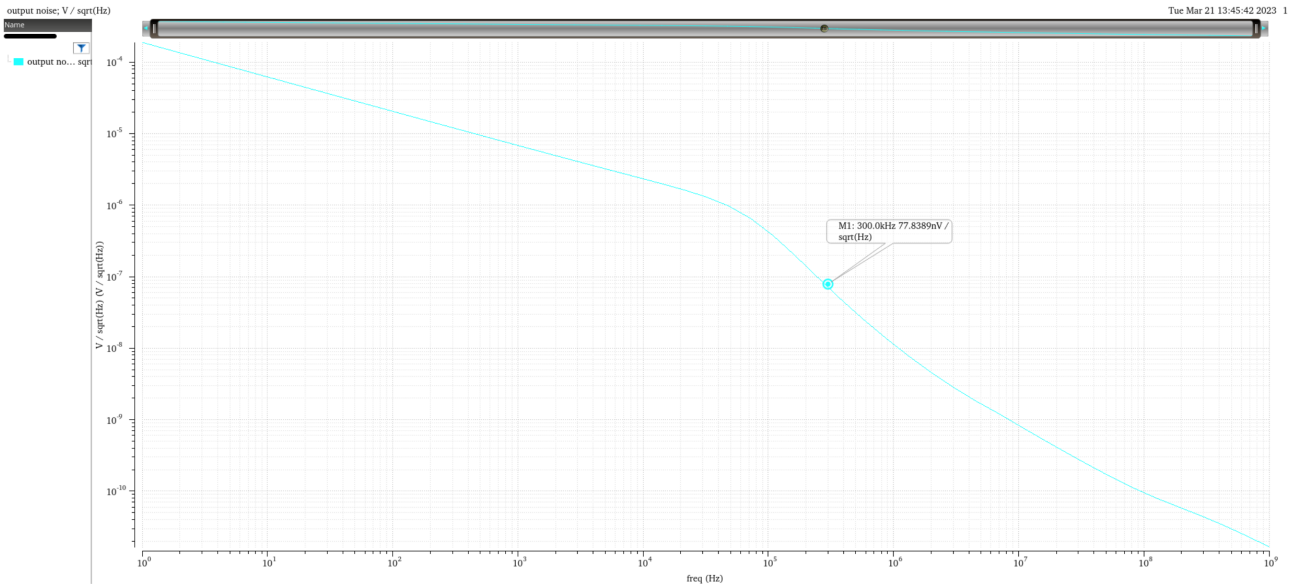

Figure 3.13: Noise spectral density without no chopping

Table 6: Main Noise contributor when chopping is applied

name of the noise source	noise contribution of total (in %)	type of noise (thermal/flicker)
R_{c1}	29.42%	thermal
PCH_{out}	4.42%	flicker
NCH_{cms1}	3.53%	flicker
NCH_{cms2}	3.53%	flicker
PCH_{in+}	2.75%	thermal
PCH_{in-}	2.75%	thermal

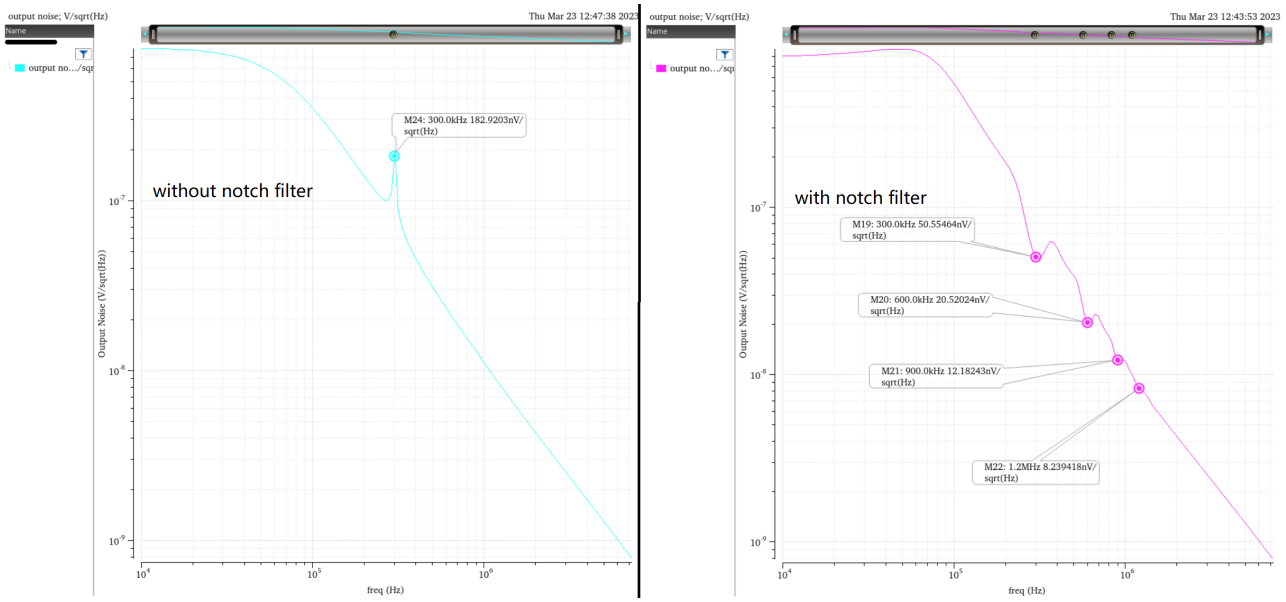

Figure 3.14: Noise spectral density with and without the notch filter

Figure 3.14 shows that the chopping frequency is higher than the corner frequency. Thus a significant attenuation in the flicker noise shall be observed.

As shown in the left figure in Figure 3.14, this is the noise spectral density when the chopping is on and the notch filter is off. The flicker noise at low frequency is shifted to the chopping frequency(f_{ch}). And in the right figure, the attenuation at $n \cdot f_{chop}$ can be observed. Note that offset can also be taken as noise at extremely low frequencies, which will be chopped and filtered as well.

The integrated noise at $27°C$ at the typical corner in $10kHz \sim 10MHz$ without chopping applied is $350\mu V_{rms}$, and with chopping applied, this value turns to $260\mu V_{rms}$. And with checking the PVT corners, the integrated noise result ranges in $134.5\mu V_{rms} \sim 507.2\mu V_{rms}$. Since the flicker noise is chopped and filtered, now the dominant noise is the thermal noise from the curvature correction resistor because the noise from the curvature correction resistor will be amplified by $\frac{R_1}{R_{1c}}$. And since the value of the feedback resistor is directly related to the bias current of the BJT($R_3 = \frac{\Delta V_{BE}}{I_{PTAT}}$). They can be attenuated by either increasing the area and bias current for BJT by the same ratio for promising the same bias density(costing more area and power) or applying bigger resistance and capacitance to the low-pass filter attached to the output(costing more area and settling time). The list of dominating noise contributors is shown in table 6.

3.3.8 PSR behavior

The PSR is promised by the capacitor attached to the gate of the output stage and the low-pass filter attached to the output. The capacitor attached to the output stage establishes AC coupling between the power supply and the gate control voltage. This helps to build an AC clean V_{gs} for the output stage. And the low-pass filter at the output helps filter the AC components out. Since $V_{out} \approx 1V$(actually slightly higher), the PSR is taken as $PSR = 20log_{10}(\frac{1}{V_{out,ac}})$. The test has been done among all PVT corners in $1Hz \sim 1GHz$; the worst case is 29dB, as shown in Figure 3.15. If the designer wants better PSR, then the capacitor in the low-pass filter attached to the output and the capacitor in the core controller could be played with.

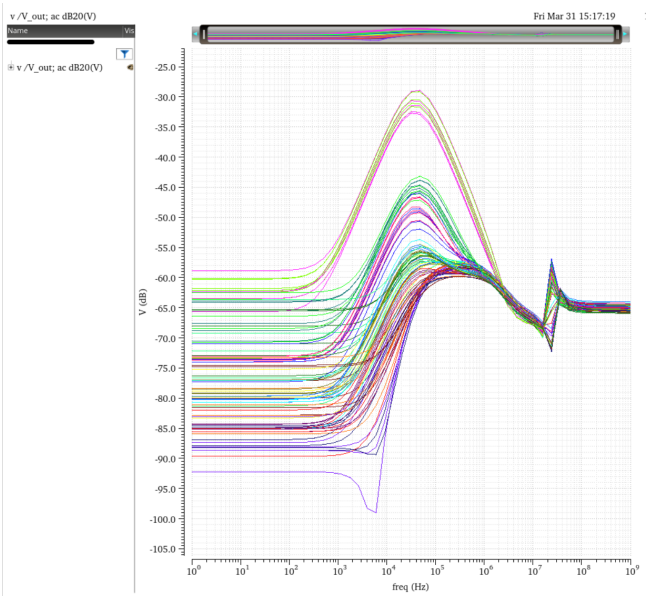

Figure 3.15: PSR among all PVT corners

3.3.9 Mismatch contribution

The aforementioned behaviors are the behaviors that do not dominate the circuit area. And the result of the previous sections proves that the mismatch and process variation from the BJT and the resistors are trimmable. Then the last step is to check the mismatch variation of each controller component and increase their area to get a low mismatch enough result, as shown in table 7. The gate area spends on each component to get an acceptable result is also in the table.

With chopping applied, the core controller does not need extra area to promise a low mismatch. It is easy to chop the mismatch to an ignorable level. This now makes the second controller the main contributor, especially the current mirror at the output stage of the second controller. At the layout level, the detailed area will be sketched graphically. In total, the mismatch contributed by the controllers which are not expected to be trimmed is $\pm 3mV$ in 5σ, which is good enough for the given project. And from the table, it can be told that if the designer wants a better mismatch behavior. What needs to be done is to increase the area for the second controller. Note that the second controller is not the main area consumer among the whole circuit, so there is still some headroom for a better mismatch result with a considerable area cost.

Table 7: Mismatch contribution of all components

name of the component	vontribution to V_{out} variation per std	gate area
total mismatch	$1.578mV$	
R and BJT	$1.44mV$	trimmable, no need more area
both controllers	$0.64mV$	$476\mu m^2$
core controller: input pair	$0.02mV$	$26\mu m^2$
core controller: current mirror (PMOS and NMOS)	$0.15mV$	$33\mu m^2$
core controller: common gate stages (PMOS and NMOS)	$0.23mV$	$78\mu m^2$
Bias circuit	$0.04mV$	$77.7\mu m^2$
Second controller: input pair	$0.266mV$	$10\mu m^2$
Second controller: current mirror	$0.206mV$	$1.77\mu m^2$
Second controller: output stage	$0.52mV$	$250\mu m^2$

Table 8: Process and mismatch variation of the V_{out}

error type	origional value per std	expect to be trimmed	post trim residual error of total (%) in 5σ
process variation	$2.14mV$	$\surd$	$\pm 0.05\%$
mismatch variation from R, BJT	$1.44mV$	$\surd$	
mismatch variation from controller	$0.64mV$	$\times$(partly)	$\pm 0.2\%$
curvature error	$1mV$	$\times$	$\pm 0.05\%$
VDD variation	$1mV$	$\times$	$\pm 0.05\%$
total			± 0.3

3.3.10 V_{out} variation

Besides the mismatch variation, the process also contributes variation to V_{out}. Table 8 shows how much each type of variation contributes to the V_{out}.

Due to the time limitation. The Monte Carlo test for all post-trim cases was not done. But the test for the worst pre-trim cases and their post-trim result is analyzed and shown in Figure 3.16.

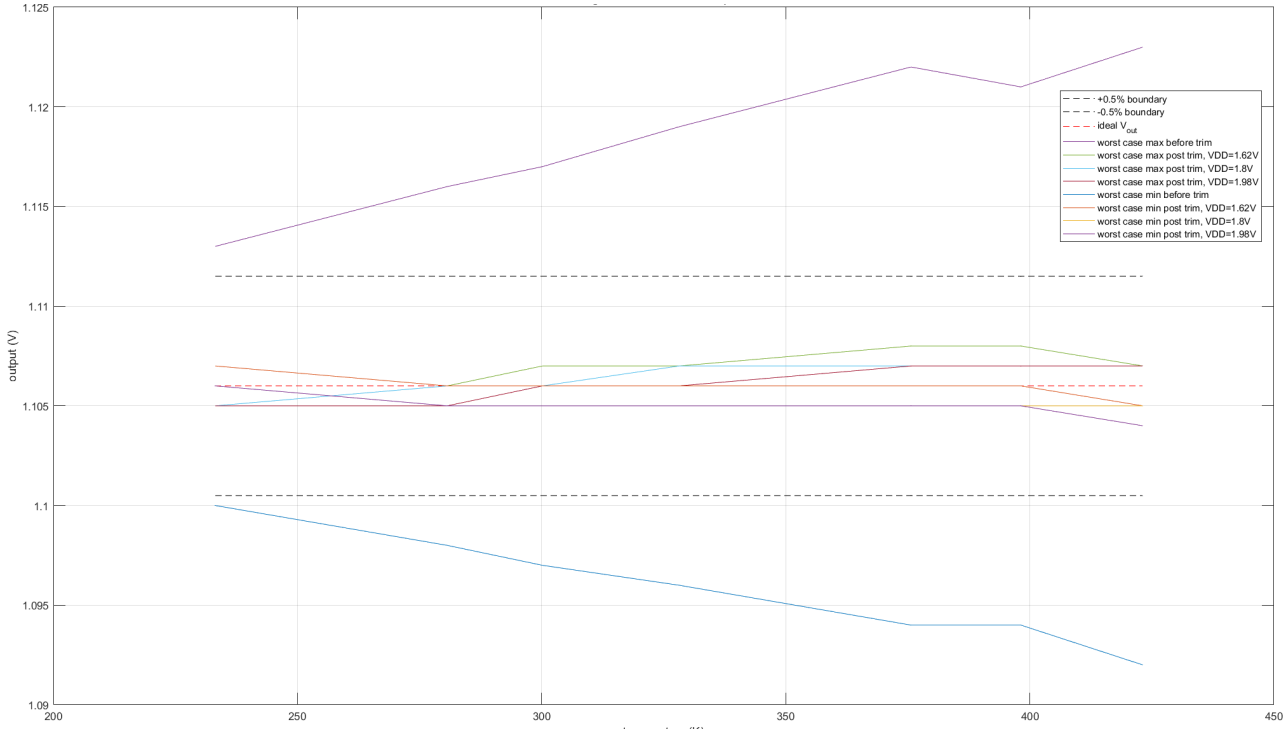

Figure 3.16: Test for the worst pre-trim cases and their post-trim result

3.3.11 Conclusion for the schematic level

The transistor-level result shows that designing such a bandgap reference circuit with acceptable power and area is possible. Even though the project requires only $\pm 0.5\%$ for accuracy, the design at the schematic level still stopped at $\pm 0.3\%$. This is because some margin for aging and nonidealities during layout must be considered.

3.4 Layout level

Layout design for 16FFC is a complex job, and since the research question for this project is to know ”if possible” and how much would be the cost and power. Considering the limited time, a general floor plan would be good enough. And for the commercial-level product, a more experienced layout engineer or several more months are needed.

3.4.1 Floor plan

The Floor plan of this project is shown in Figure 3.17. As mentioned in the beginning, the most area-consuming components are the capacitors used for the low-pass filters. Then comes the BJTs, then the current mirror in the second controller. What is different from most circuit designs is that normally the input pairs of the controllers are the main area consumers, but in this design, they are designed to be very small and with more mismatch than in the normal case. And saved spaces are used for the capacitor to minimize the effective mismatch. Besides the input pairs, using current mirrors in the signal path is also been avoided as much as possible. The Figure shows that even though the current mirror at the second controller’s output stage is not the main area-consumer, it’s still preferred to be used less. There was an old design based on C040GF that only achieved 1.25% without chopping and

Table 9: Area and current consumed by each component

component	area	current	comment
core controller capacitor	$0.01mm^2$		cancel input offset
LPF capacitor	$0.005mm^2$		due to the thermal noise, if raise BJT size by 4 and bias current by 4 the size of LPF can be 2 times less
core controller(except the capacitor)	$0.0024mm^2$	$30 \sim 60\mu A$	for a higher speed controller works properly at chopping frequency
current mirror at output stage of second controller	$0.0012mm^2$		current mirror in signal path need a very low mismatch if no extra mismatch cancelation techniques are applied
second controller(except output stage)	$0.0012mm^2$	$10 \sim 20\mu A$	less speed and msimatch requirement
bias circuit	$0.0012mm^2$	$5 \sim 10\mu A$	in case of unwwanted steady states among 5σ cases.
core BJT	$0.0013mm^2$	$4 \sim 7\mu A$	due to the shot noise of BJT, the bias current for BJT needs to be big enough
curvature correction BJT	$0.006mm^2$	$16\mu A$	the constant biasing current need to be much bigger than the extra current flows in R_{c1}
total	$0.0263mm^2$	$55 \sim 110\mu A$	
for comparasion: old design(accuracy $\pm 1.25\%$)	$0.02675mm^2$	$4.5 \sim 13.1\mu A$	

only decreased mismatch with more area. Table 9 shows the area consumed and quiescent current consumed by each component($VDD = 1.8V$).

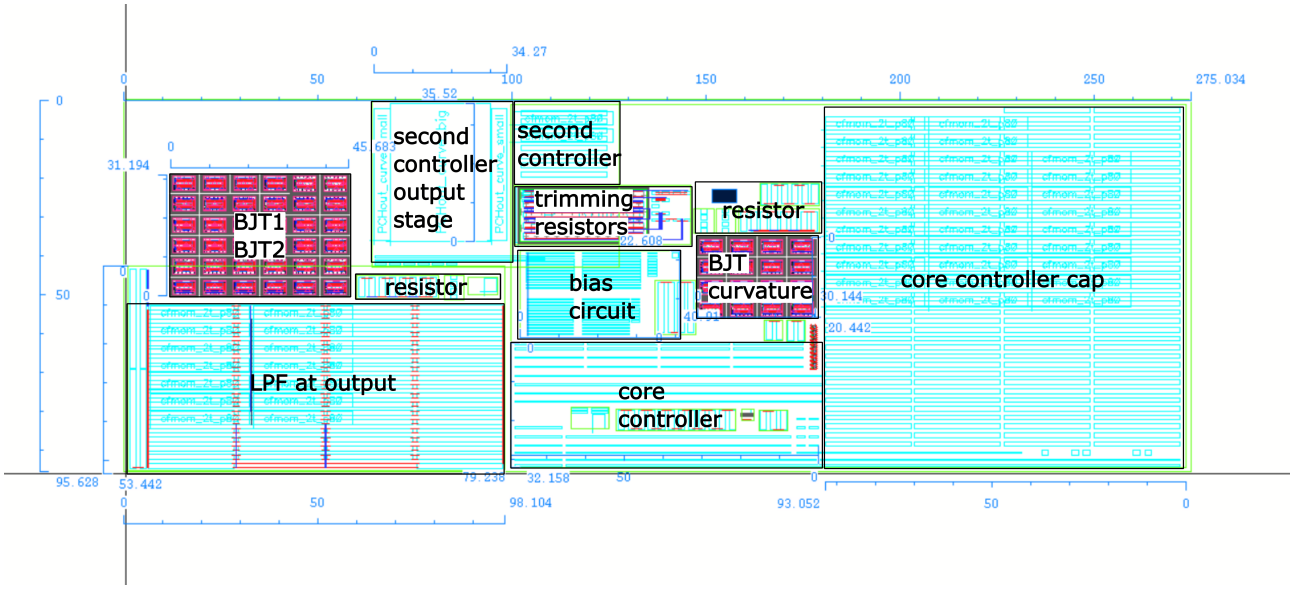

Figure 3.17: Floor plan

3.4.2 Conclusion for layout level

This project leads to a circuit with chopping applied whose area is even slightly less than the old version. This is because the old design only solved the mismatch by one aspect: area. But this project solves the mismatch by area and frequency domain. On the other hand, this project's result requires 8.5 times in power for a higher speed core controller and curvature correction part. Hopefully, this is still in the acceptable range.

Table 10: Result list for the project

requirement	value	hard or soft?	result value
5σ inaccuracy	$< \pm 0.5\%$	hard	$\pm 0.3\%$
temperature range	$-40°C \sim 150°C$	hard	
supply voltage range	$1.62V \sim 1.98V$	hard	
output integrated noise($10kHz \sim 10MHz$)	$< 500\mu V_{rms}$	soft	$< 507.2\mu V_{rms}$
PSR($20log(\frac{V_{ref}}{V_{out,AC}})$ in $1Hz \sim 1GHz$)	$> 25dB$	soft	$> 29dB$
power	$< 0.18mW$	soft	$< 0.198mV$
area	$< 0.3mm^2$	soft	$0.263mm^2$
the temperature at which trimming is performed	$125°C$	hard	$125°C$
settling time	$< 200\mu s$	soft	$< 160\mu s$

4 Results

This section is for summarizing the result of the project and having some comments on the result. Note that the simulation result is only based on the schematic-level design.

4.1 Requirement list

The report started with a requirement list. And now, it's time to see if the result meets the requirement, as shown in Table 10.

4.1.1 Accuracy

The accuracy at the schematic level shall be higher than the requirement in case of aging and nonidealities from the layout. If more accuracy is required, follow the suggestions in section 3.3.9.

4.1.2 Noise

The noise reached the requirement. And it is possible if the designer needs the circuit to perform less noise. Follow the suggestions in section 3.3.7.

4.1.3 PSR

The PSR is higher than the requirement. Since the PSR is not the project's key requirement, the design flow did not start with that. And if the designer wants to arrange the PSR, please follow the suggestions in section 3.3.8

4.1.4 Power

The power consumed by this project is much more than the old design. After discussing with the company, this is an acceptable value. If the designer wants to save some more power, please start by shrinking the current budget of the core controller.

4.1.5 Area

The area was the key aspect during the design and the current design already leads to the most area-saving result. The area may be slightly optimized during the layout design phase.

4.1.6 Temperature at which trimming is performed

It is always better to trim at the midpoint between $-40°C$ and $150°C$, which is $55°C$. Because this can split the temperature-independent error equally on both sides of ideal V_{out}.

4.1.7 Settling time

The settling time meets the requirement. If the designer wants to redesign the settling time, please follow the suggestions in section 3.3.6.

5 Conclusion

The conclusion of this project can be split into two parts. The answer to the research question and what can be done for this project in the future.

5.1 Answer to the research question

To remind the reader, the research question is repeated again.

5.1.1 Research question

Is it possible to design a bandgap reference voltage source that achieves 5-sigma Inaccuracy of $\pm 0.5\%$ from -40C to 150C under the C040GF or 16FFC process? If possible, how much is the cost of the area and power?

5.1.2 Answer

It is possible to design a bandgap reference voltage source that achieves 5-sigma Inaccuracy of $\pm 0.5\%$ from -40C to 150C under the C040GF and 16FFC process. And a design in 16FFC shows that the cost of power will increase to an acceptable value but slightly higher than most designs because of the power consumed by the controller for chopping and curvature correction to reach high accuracy. The design also shows that there will be no extra area cost for such a high requirement. And with a proper design, it is possible to reach a smaller area with chopping.

5.2 Future Work

Due to the complexity of 16FFC and the limitation of time. It is quite a pity that this project can not have a layout ready for tape out. But the result so far is still quite surprising, especially the area result. This design still points to a promising integrated bandgap reference circuit that is worth trying in layout and tape out. The next step could be extending the design to the layout level and checking if the circuit could perform wanted V_{out}. There is a fat chance that this could be the answer to the next generation of bandgap voltage reference circuits in in-vehicle networks.

Bibliography

[1] A. van Staveren, C. Verhoeven, and A. van Roermund, "The design of low-noise bandgap references," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 43, no. 4, pp. 290–300, 1996.

[2] G. Ge, C. Zhang, G. Hoogzaad, and K. A. A. Makinwa, "A single-trim cmos bandgap reference with a 3σ inaccuracy of $\pm 0.15 - 40^{\circ}$c to 125°c," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp. 2693–2701, 2011.

[3] R. T. Perry, S. H. Lewis, A. P. Brokaw, and T. R. Viswanathan, "A 1.4 v supply cmos fractional bandgap reference," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2180–2186, 2007.

[4] M. Stefan and T. O'Dwyer, "Curvature correction method for a bandgap voltage reference," in *IET Irish Signals and Systems Conference (ISSC 2008)*, pp. 134–137, 2008.

[5] P. Malcovati, F. Maloberti, C. Fiocchi, and M. Pruzzi, "Curvature-compensated bicmos bandgap with 1-v supply voltage," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 7, pp. 1076–1081, 2001.