Bandgap Reference Design at the 14-Nanometer FinFET Node

Lucas J. Prilenski
ljp2706@rit.edu

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation

Accessed from
Dedication

This thesis is dedicated to my parents and my fiancé.
Acknowledgments

I am grateful for the encouragement of my parents; without them, I know I would not be where I am today. My fiancé Kathryn provided me with encouragement over the years and has strengthened my dedication towards analog design with her never ending words of motivation. I would like to thank my committee members Dr. Moon, Dr. Mukund, and Professor Indovina for the effort that they put into helping me with my thesis. I chose Dr. Moon, Dr. Mukund, and Professor Indovina as my committee members because they have truly inspired me to aspire towards my goals of proficiency in analog/mixed signal design. I would like to thank Dr. Mukund for giving me the amazing opportunity to work with him on a project that I was very passionate about; I have learned so much working with you. And lastly, I would like to thank all of those that I have met along this journey; your support has made all the difference to me.
Abstract

Bandgap Reference Design at the 14-Nanometer FinFET Node

Lucas J. Prilenski

Supervising Professor: Dr. P. R. Mukund

As supply voltages continue to decrease, it becomes harder to ensure that the voltage drop across a diode-connected BJT is sufficient to conduct current without sacrificing die area. One such solution to this potential problem is the diode-connected MOSFET operating in weak inversion. In addition to conducting appreciable current at voltages significantly lower than the power supply, the diode-connected MOSFET reduces the total area for the bandgap implementation. Reference voltage variations across Monte Carlo perturbations are more pronounced as the variation of process parameters are exponentially affected in subthreshold conduction. In order for this proposed solution to be feasible, a design methodology was introduced to mitigate the effects of process variation. A 14 nm bandgap reference was created and simulated across Monte Carlo perturbations for 100 runs at nominal supply voltage and 10% variation of the power supply in either direction. The best case reference voltage was found and used to verify the proposed resistive network solution. The average temperature coefficient was measured to be 66.46 $\frac{ppm}{^\circ C}$ and the voltage adjustment range was found to be 204.1 mV. The two FinFET subthreshold diodes consume approximately 2.8% of the area of the BJT diode equivalent. Utilizing an appropriate process control technique, subthreshold bandgap references have the potential to overtake traditional BJT-based bandgap architectures in low-power, limited-area applications.
Commonly Used Symbols and Abbreviations

BJT Bipolar Junction Transistor
$C_{ox}$ Oxide Capacitance
CLM Channel Length Modulation
CM Current Mirror
CMOS Complementary Metal Oxide Semiconductor
CTAT Complimentary to Absolute Temperature
DVECM Drain Voltage Equalization Current Mirror
DIBL Drain-Induced Barrier Lowering
eV Electron Volts
$\gamma$ Body Effect Parameter
$g_m$ Transconductance
$I_D$ Drain Current
$k$ Boltzmann’s Constant
$k'$ $\mu C_{ox}$
L Physical Channel Length
$\mu$ Mobility
MOSFET Metal Oxide Semiconductor Field-Effect Transistor
$\phi_t$ Thermal Voltage
PTAT Proportional to Absolute Temperature
RFID Radio-Frequency Identification
T Temperature
TC Temperature Coefficient
$V_{GS}$ Gate-to-Source Voltage
$V_{OV}$ Overdrive Voltage
$V_T$ Threshold Voltage
W Physical Channel Width
## Contents

Dedication ........................................................................................................ iii

Acknowledgments ............................................................................................ iv

Abstract ........................................................................................................... v

Commonly Used Symbols and Abbreviations .................................................. vi

1 Introduction .................................................................................................. 1

2 Basic Semiconductor Physics Background ................................................. 3
   2.1 Current Density ..................................................................................... 3
   2.2 Mobility .............................................................................................. 5
   2.3 The Abrupt pn Junction Diode ............................................................ 5
   2.4 Bipolar Junction Transistors ............................................................... 6
   2.5 MOSFET Device Structure ................................................................. 6
   2.6 FinFET Device Structure .................................................................... 7
   2.7 Regions of Operation ......................................................................... 8
      2.7.1 Weak Inversion ........................................................................... 8
      2.7.2 Moderate Inversion .................................................................... 8
      2.7.3 Strong Inversion ......................................................................... 9
   2.8 Concluding Remarks .......................................................................... 9

3 Introduction of the FinFET ........................................................................ 10
   3.1 FinFET Device ................................................................................... 10
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>Temperature Dependency</td>
<td>46</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Temperature Dependence of Resistors</td>
<td>46</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Temperature Dependence of Field Effect Transistors</td>
<td>47</td>
</tr>
<tr>
<td>5.4.3</td>
<td>PTAT Generation</td>
<td>47</td>
</tr>
<tr>
<td>5.4.4</td>
<td>CTAT Generation</td>
<td>48</td>
</tr>
<tr>
<td>5.5</td>
<td>Device Sizing Considerations</td>
<td>48</td>
</tr>
<tr>
<td>5.6</td>
<td>Drain Voltage Equalization Current Mirror Bandgap</td>
<td>49</td>
</tr>
<tr>
<td>5.7</td>
<td>Sensitivity Issues</td>
<td>57</td>
</tr>
<tr>
<td>5.8</td>
<td>Modified FinFET Drain Voltage Equalization CM Bandgap</td>
<td>58</td>
</tr>
<tr>
<td>5.9</td>
<td>DC Biasing Considerations</td>
<td>61</td>
</tr>
<tr>
<td>5.10</td>
<td>Resistive Trimming Networks</td>
<td>62</td>
</tr>
<tr>
<td>5.11</td>
<td>Applications</td>
<td>65</td>
</tr>
<tr>
<td>5.12</td>
<td>Matching</td>
<td>66</td>
</tr>
<tr>
<td>5.13</td>
<td>Higher Order Curvature Correction</td>
<td>68</td>
</tr>
<tr>
<td>5.14</td>
<td>Generalized Bandgap Design Methodology</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>Bandgap Reference Results</td>
<td>72</td>
</tr>
<tr>
<td>6.1</td>
<td>14-nm Drain Voltage Equalization CM Bandgap</td>
<td>72</td>
</tr>
<tr>
<td>6.1.1</td>
<td>PVT Results</td>
<td>72</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Monte Carlo Results</td>
<td>74</td>
</tr>
<tr>
<td>6.1.3</td>
<td>$V_{DD}$ vs. $V_{ref}$</td>
<td>76</td>
</tr>
<tr>
<td>6.1.4</td>
<td>PSR Graph</td>
<td>78</td>
</tr>
<tr>
<td>6.1.5</td>
<td>Transient Analysis</td>
<td>80</td>
</tr>
<tr>
<td>6.1.6</td>
<td>Trim Results</td>
<td>82</td>
</tr>
<tr>
<td>6.2</td>
<td>Higher Order Compensation</td>
<td>84</td>
</tr>
<tr>
<td>6.3</td>
<td>45 nm Drain Voltage Equalization CM Bandgap</td>
<td>86</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Monte Carlo Results</td>
<td>87</td>
</tr>
<tr>
<td>6.3.2</td>
<td>$V_{DD}$ vs. $V_{ref}$</td>
<td>88</td>
</tr>
<tr>
<td>6.3.3</td>
<td>PSR Graph</td>
<td>89</td>
</tr>
</tbody>
</table>
List of Tables

6.1 Bandgap Results and Comparison . . . . . . . . . . . . . . 93
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Planar CMOS Cross-section</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Planar CMOS Device Structure</td>
<td>7</td>
</tr>
<tr>
<td>2.3</td>
<td>FinFET 3D Device Structure</td>
<td>8</td>
</tr>
<tr>
<td>4.1</td>
<td>Amplifier Block Diagram</td>
<td>28</td>
</tr>
<tr>
<td>4.2</td>
<td>Class A Output</td>
<td>29</td>
</tr>
<tr>
<td>4.3</td>
<td>Effective Load Simulation</td>
<td>30</td>
</tr>
<tr>
<td>4.4</td>
<td>Brokaw Cell Schematic</td>
<td>32</td>
</tr>
<tr>
<td>4.5</td>
<td>Differential Input Helper Transistors</td>
<td>35</td>
</tr>
<tr>
<td>4.6</td>
<td>Monticelli Bias Schematic [1]</td>
<td>36</td>
</tr>
<tr>
<td>4.7</td>
<td>Two Stage Op-Amp Schematic With Helper Device</td>
<td>37</td>
</tr>
<tr>
<td>5.1</td>
<td>Temperature Dependency Superposition</td>
<td>40</td>
</tr>
<tr>
<td>5.2</td>
<td>Implemented 14 nm Bandgap Reference</td>
<td>41</td>
</tr>
<tr>
<td>5.3</td>
<td>MOS Diode Voltage Temperature Coefficient Simulation</td>
<td>42</td>
</tr>
<tr>
<td>5.4</td>
<td>NMOS Cascode Current Mirror</td>
<td>43</td>
</tr>
<tr>
<td>5.5</td>
<td>Simple Current Mirror Schematic</td>
<td>46</td>
</tr>
<tr>
<td>5.6</td>
<td>Beta Bandgap Reference Schematic</td>
<td>49</td>
</tr>
<tr>
<td>5.7</td>
<td>Bootstrapped Widlar Current Source</td>
<td>54</td>
</tr>
<tr>
<td>5.8</td>
<td>Inverter Based Start-Up</td>
<td>55</td>
</tr>
<tr>
<td>5.9</td>
<td>Leakage Start-Up Method</td>
<td>56</td>
</tr>
<tr>
<td>5.10</td>
<td>Dynamic Start-Up Circuit</td>
<td>56</td>
</tr>
<tr>
<td>5.11</td>
<td>14 nm Bandgap Reference with an Internally Biased Amplifier</td>
<td>58</td>
</tr>
<tr>
<td>5.12</td>
<td>Series Trim Network Schematic</td>
<td>63</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.13</td>
<td>Parallel Trim Network Schematic</td>
<td>64</td>
</tr>
<tr>
<td>5.14</td>
<td>Effective Trim Network Schematic</td>
<td>64</td>
</tr>
<tr>
<td>5.15</td>
<td>Precision Current Reference</td>
<td>65</td>
</tr>
<tr>
<td>5.16</td>
<td>Higher Order Curvature Correction Schematic</td>
<td>68</td>
</tr>
<tr>
<td>5.17</td>
<td>Complete Bandgap Schematic</td>
<td>71</td>
</tr>
<tr>
<td>6.1</td>
<td>14 nm Reference Voltage PVT Variations</td>
<td>73</td>
</tr>
<tr>
<td>6.2</td>
<td>14 nm Reference Voltage Monte Carlo Analysis</td>
<td>75</td>
</tr>
<tr>
<td>6.3</td>
<td>14 nm Voltage Headroom with the Resistive Trim Network</td>
<td>77</td>
</tr>
<tr>
<td>6.4</td>
<td>14 nm Voltage Headroom with the MOS Trim Network</td>
<td>78</td>
</tr>
<tr>
<td>6.5</td>
<td>14 nm PSR Graph</td>
<td>79</td>
</tr>
<tr>
<td>6.6</td>
<td>14 nm Transient Start-Up Analysis</td>
<td>81</td>
</tr>
<tr>
<td>6.7</td>
<td>14 nm Digital Trim Network Switching</td>
<td>82</td>
</tr>
<tr>
<td>6.8</td>
<td>14 nm 6-Bit Resistive Trim Network</td>
<td>83</td>
</tr>
<tr>
<td>6.9</td>
<td>14 nm 6-Bit MOS Trim Network</td>
<td>84</td>
</tr>
<tr>
<td>6.10</td>
<td>14 nm Compensation Comparison</td>
<td>85</td>
</tr>
<tr>
<td>6.11</td>
<td>45 nm Reference Voltage PVT Variations</td>
<td>87</td>
</tr>
<tr>
<td>6.12</td>
<td>45 nm Reference Voltage Monte Carlo Analysis</td>
<td>88</td>
</tr>
<tr>
<td>6.13</td>
<td>45 nm Voltage Headroom with the Resistive Trim Network</td>
<td>89</td>
</tr>
<tr>
<td>6.14</td>
<td>45 nm PSR Graph</td>
<td>89</td>
</tr>
<tr>
<td>6.15</td>
<td>45 nm Transient Start-Up Analysis</td>
<td>90</td>
</tr>
<tr>
<td>6.16</td>
<td>45 nm 6-Bit Resistive Trim Network</td>
<td>91</td>
</tr>
<tr>
<td>6.17</td>
<td>Layout Comparisons</td>
<td>92</td>
</tr>
<tr>
<td>A.1</td>
<td>Implemented 14 nm Bandgap Reference</td>
<td>100</td>
</tr>
<tr>
<td>C.1</td>
<td>Brokaw Cell Schematic</td>
<td>106</td>
</tr>
<tr>
<td>D.1</td>
<td>Higher Order Curvature Correction Schematic</td>
<td>109</td>
</tr>
<tr>
<td>E.1</td>
<td>Two Stage Op-Amp Schematic With Helper Device M₈</td>
<td>111</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Supply voltages continue to decrease as minimum feature sizes of transistors scale down. It becomes harder to ensure that the voltage drop across a diode BJT is sufficient to conduct current without sacrificing die area and without experiencing the effects of voltage headroom limitations. One such solution to this problem is the diode-connected MOSFET operating in weak inversion.

This thesis discusses an implementation of a sub-one-volt bandgap reference in a 14-nanometer FinFET process. The design was a modified version of the Banba bandgap circuit [2]. The primary difference was the use of subthreshold diode-connected FinFET PMOS devices. The familiar PTAT expression derived for bipolar transistors was also applicable to FinFETs in the subthreshold regime, as the gate-to-source voltage exhibits an exponential dependence much like a BJT as opposed to the linear-to-squared dependence when in saturation.

This thesis details the use of MOS diodes within a bandgap reference that typically used BJT diodes. Chapter 2 provides the necessary background on
semiconductor device physics. Chapter 3 introduces the device structure used in this work, the FinFET. Chapter 4 discusses the error amplifier used within the bandgap, the design modifications, and how the op amp impacts the overall design. Chapter 5 provides an overview of how a bandgap reference operates, explains sizing considerations, and discusses the common non-idealities of bandgap references. Chapter 6 shows the results of the 14 nm bandgap and compares it to other CMOS bandgaps as well as one designed by the author at the 45 nm node. Chapter 7 draws important conclusions about the results of this work.
Chapter 2

Basic Semiconductor Physics Background

A strong foundation in the knowledge of semiconductor device physics is crucial in analyzing the operation of a bandgap reference. The fundamentals of device physics can be used to prove that to a first-order approximation, FinFET devices within weak inversion resemble the BJT device.

2.1 Current Density

There are two main components that comprise the total current in a carrier injection device such as a BJT, or in a MOS diode, and they are drift and diffusion current. Diffusion current is dependent on the gradient of the doping profile. Equations 2.1 and 2.2 below show the diffusion terms of the current density equation for electrons and holes.

\[ J_{\text{diffusion}, n} = qD_n \frac{dn(x)}{dx} \]  \hspace{1cm} (2.1)

\[ J_{\text{diffusion}, p} = -qD_p \frac{dp(x)}{dx} \]  \hspace{1cm} (2.2)
where \( D \) is related to the mobility, \( \mu \), as in the Einstein relation shown below.

\[
D = \mu \left( \frac{kT}{q} \right) \quad (2.3)
\]

Drift current has a dependence on electric field: the velocity of the carriers is dictated by mobility and the electric field. The drift current density components for electrons and holes are given by equations 2.4 and 2.5 below.

\[
J_{\text{drift},n} = qn(x)\mu_n \mathcal{E}(x) \quad (2.4)
\]

\[
J_{\text{drift},p} = qp(x)\mu_p \mathcal{E}(x) \quad (2.5)
\]

For a carrier-injection device such as a BJT, the diffusion term comprises most of the total current density, whereas in a field-effect device such as a MOSFET, typically it is dominated by the drift component due to the high lateral fields. Operation of a MOSFET in the subthreshold region implies low field and the MOSFET behaves similar to a BJT—i.e., it acts as a diffusion-based device.


2.2 Mobility

One of the key parameters dictating the drain current of a MOSFET is the carrier mobility. There are two main components of mobility. At high temperatures phonon scattering dominates due to lattice vibration as a result of thermal excitation, and at low temperatures ionized impurity scattering dominates which occurs when impurity charge exerts an attractive or repulsive influence on the carrier within the channel of the device.

2.3 The Abrupt pn Junction Diode

The built-in potential that must be overcome to conduct appreciable current is given by equation 2.6 for a non-degenerately doped junction.

\[
\psi_{bi} = \frac{kT}{q} \ln \left( \frac{N_A N_D}{n_i^2} \right)
\]  

(2.6)

The value of the built-in potential typically ranges from 0.6 to 0.8 electron volts (eV). As power supplies approach these values, it becomes harder to keep current mirrors in saturation as the diode requires most of the voltage across it to conduct current. One solution that could be implemented would be to increase the size of the device. However, this is impractical as it would consume a great deal of space. This is the motivation for investigating diodes with lower potential barriers.
2.4 Bipolar Junction Transistors

Vertical PNP bipolar junction transistors can be used within a voltage reference circuit; however the area of the device is very large in order to obtain a small base resistance. These devices also cannot be cascoded as a result of the characteristic that the collectors of vertical PNP devices are grounded.

2.5 MOSFET Device Structure

Figure 2.1 shows a basic CMOS process. P-type substrates are typically used as they are cheaper to fabricate. NMOS $n^+$ implants for the source and drain do not require a well because the substrate acts as the well. The body contact is a $p^+$ implant since a direct metal to low-doped semiconductor connection will form a non-ohmic Schottky contact. The oxide layer separates the gate from the channel, which is defined to be the region between the source and drain underneath the gate. The PMOS device is similar to the NMOS device but it must be placed within an n-well due to the $p^+$ source and drain implants.
Figure 2.1: Planar CMOS Cross-section

Figure 2.2 shows the traditional planar CMOS process. It can be seen that the source and drain are planarized with the gate directly above.

2.6 FinFET Device Structure

The FinFET architecture obtains its name from the fin-like appearance of the raised source-drain structure, which can be seen in Figure 2.3. The source-drain channel extends into the vertical plane, which allows for a higher degree of control over the channel.
2.7 Regions of Operation

The three main regions of operation of the metal-oxide-semiconductor field effect transistor (MOSFET) are subthreshold, linear, and saturation.

2.7.1 Weak Inversion

Subthreshold conduction occurs when the voltage applied across the gate and source nodes of the device is significantly less than the threshold voltage. In this region of operation very little current flows, which is suitable for low power applications. This region of operation is primarily diffusion-based.

2.7.2 Moderate Inversion

The second region of operation, the linear region, occurs when the gate-to-source voltage is only several thermal voltages less than or greater than the threshold voltage. This region of operation is extremely complex to
properly model as not many simplifications can be made to the charge-based
equations governing the current flow within the channel of the device. The
moderate inversion region of operation is a combination of drift and diffu-
sion current.

2.7.3 Strong Inversion

The last, and perhaps most often used region of operation, is strong
inversion-saturation. This occurs when the gate-to-source voltage is sig-
nificantly greater than the thermal voltage and the drain-to-source voltage
exceeds the potential required to saturate the velocity of the carriers within
the channel. Current in the strong inversion region of operation is mainly
from drift current.

2.8 Concluding Remarks

This work makes use of MOSFET devices operating in weak inversion
as the current density is primarily controlled by the diffusion current density
component. This results in MOS devices that operate similarly to BJT de-
vices; there is an exponential relationship between current and controlling
gate voltage. The key difference between these two devices is the fact that
the MOS device does not have a built-in potential that must be overcome
to conduct current, which reduces the minimum operating voltage of the
device.
Chapter 3

Introduction of the FinFET

3.1 FinFET Device

FinFET devices are categorized as vertical double-gate devices which can be modeled as two SOI devices with the sources connected and the drains connected. There are two different configurations: the three-terminal device in which both gates are connected together and the four-terminal device where the front gate is used as the controlling gate and the back gate provides a fixed bias [3]. Shorting the front and back gates together yields a higher degree of control over the channel whereas biasing the back gate independently of the front gate allows for threshold voltage adjustments. The significance of an independent back gate is that it gives a way for back-gate control circuitry to mitigate the effects of process, temperature, and voltage variations.
3.2 Benefits

Benefits of the FinFET structure include a lower drain-to-source conductance which allows for a much higher output impedance, leading to improvements in gain. The subthreshold slope is almost ideal, the body effect is almost eliminated, there are small parasitic capacitances, and the devices exhibit very high transconductance [4]. The numerous benefits of FinFET devices mitigate the short-channel effects (SCEs) that arise from scaling down the minimum gate length, which promotes the architecture as one of the most promising technologies.

3.3 Short-Channel Effects

Several short-channel effects which are present in FinFET devices as well as bulk CMOS devices are channel-length modulation (CLM), drain-induced barrier lowering (DIBL), punch through, mobility degradation, and increased subthreshold swing. CLM is more pronounced in short-channel devices [5]; the electrical length of the channel is effectively reduced as the length $\Delta L$ of the velocity saturation region increases with $V_{DS}$. This gives rise to a linear increase in drain current in the saturation region. DIBL causes variations in threshold voltage and consequentially, the drain current [5], due to the impact of the drain voltage control on the inversion channel. Punch through is an extreme case of CLM or DIBL[5]. It occurs when the depletion regions of the drain and the source become large.
enough that they merge into a single depletion region. This undesirable effect causes large current variations with slight changes in drain-to-source potential variations. Punch through is a key factor in determining the limits of the maximum operating voltage of the device [5]. Mobility in the inversion layer is significantly lower than the bulk mobility [5]. Electrons towards the surface of the device shield electrons further down in the device from the roughness of the surface and the strong electric field. Mobility is lowest at the surface due to mobility degradation in high vertical fields. At high longitudinal electric fields, the linear approximation between electric field and velocity no longer holds (the proportionality factor being mobility) [5]. Eventually the velocity of the electrons saturates at a given value dependent on the material of the device. Lastly, subthreshold swing is defined to be the change in voltage required to change the magnitude of the drain current by a factor of ten. As technology scales down, devices tend to leak more, which degrades the subthreshold swing. It becomes harder to control the channel, and consequentially, it becomes harder to shut off the device.

### 3.3.1 Comments on Short-Channel Effects

In double-gate MOSFETs, the gates control the energy barrier between the source and the drain, which allows for manipulation of the barrier independently of channel doping concentration [6]. Reducing SCEs without introducing impurities into the channel will reduce the effects of statistical
dopant fluctuations. SCEs cause an increase in subthreshold swing with decreasing channel length [6]. Even with a relatively large subthreshold swing, the devices do not exhibit excessive subthreshold leakage given proper sizing of the fin width. As fin width decreases, the subthreshold swing also decreases. A low value for subthreshold swing gives better control over the channel and lower values for subthreshold leakage.

Devices with increasing silicon fin widths exhibit an increase in drain current [6]. The threshold voltage of the device decreases with increasing drain-to-source voltage; this effect is DIBL. Creating a silicon fin width that is slightly less than the length of the channel will reduce the effects of drain-induced barrier lowering [6]. As gate length decreases, the device threshold voltage quasi-linearly decreases until a critical gate length value where the threshold voltage rolls off [6] and begins to rapidly decrease.

Choosing the proper gate work function [6] is essential as control over the threshold voltage through body doping is not ideal for vertical double-gate devices because the process is designed to suppress the effects of the floating body, improve mobility, and reduce the impact of random dopant fluctuations. Double-gate devices must take care to ensure proper alignment of the gates to minimize parasitic capacitances and resistances, as well as to maximize control over the channel [6].

The back gate controls several parameters such as threshold voltage, subthreshold swing, and the short-channel effect—DIBL [7]. The front gate of the device also introduces shifts in subthreshold swing and DIBL due to
the effects of the transverse electric field. The transverse electric field approaches zero when the gate-to-source voltages of the front and back gates are equal.

In the symmetric double-gate device, the high transverse electric field reduces the impact of DIBL by moving the leakage currents to the front surface rather than the back surface [7] which is typical of a fully depleted SOI device. When the front gate-to-source voltage is not equal to the back gate-to-source voltage, the threshold voltage increases, which changes the sub-threshold characteristics. Choosing gate work functions that are not equal will also change the transverse electric field and as a result, influence the SCEs [7]. Decreasing the front gate-to-source voltage tends to increase the control of the SCEs to the point where they are significantly better than in the case where the transverse electric field is zero. Decreasing the front gate-to-source voltage is a viable method for reducing SCEs; however, the drawback is carrier mobility degradation due to the high transverse field [7] present and as a consequence, current drive is reduced.

3.4 Derivation of FinFET Drain Current Equations

In an undoped silicon film, the potential of any point in the device is satisfied by Poisson’s equation [8] given in equation 3.1 below, where $V$ is
given as the quasi-fermi potential.

\[ \frac{d^2\psi}{dx^2} = \frac{q}{\epsilon_{si}} n_i e^{\frac{q(\psi - V)}{kT}} \]  

(3.1)

The ability to calculate surface potential in the channel of a four-terminal device is crucial, since surface potential dictates the strength of the inversion layer—the stronger the inversion layer, the more inversion charge is present. If a lateral electric field is present, more current flows as inversion strength increases.

A simplification used in deriving the equations for the three modes of operation was the gradual channel approximation [9]. The source-to-body voltage is zero and the threshold voltage is constant along the channel of the device. Using the continuous analytic solution [9], equations for each region of operation can be derived. That is, by applying the appropriate boundary conditions, the drain current equation for linear, saturation, and subthreshold can be realized. Integrating the Poisson equation twice [9] yields expression 3.2 for the potential in the channel where \( \beta \) is a parameter derived from a boundary condition.

\[ \psi(x) = V - \frac{2kT}{q} \ln \left( \frac{t_{si}}{2\beta} \sqrt{\frac{q^2n_i}{2\epsilon_{si}kT}} \cos \left( \frac{2\beta x}{t_{si}} \right) \right) \]  

(3.2)

Pao-Sah’s integral (3.3) can be used to determine the drain current for the
symmetric double gate device. The bounds of integration reflect the potentials at the source and drain ends of the channel.

\[ I_{ds} = \mu \frac{W}{L} \int_0^{V_{DS}} (-Q_i(V)) \, dV \]  

(3.3)

The inversion charge is the total mobile charge per unit gate area within the channel of the device. The integration starts at the source end of the channel where typically, the potential is zero and the upper bound of the integration is at the drain end of the channel which is a \( V_{DS} \) potential increase across the length of the channel. The Pao-Sah integral can be solved for the subthreshold, linear, and saturation regions of operation. A generalized equation can be obtained and the appropriate boundary conditions can be applied for each region of operation in order to determine the drain current equation for a symmetric double-gate MOSFET.

The drain current expression in saturation is defined by equation 3.4 which was derived using the continuous analytic model [9].

\[ I_{ds,\text{sat}} = \mu C_{ox} \frac{W}{L} \left[ (V_G - V_T)^2 - \frac{8rk^2T^2}{q^2}e^{\frac{q(V_G - V_0 - V_{DS})}{kT}} \right] \]  

(3.4)

\( r \) is defined as a structural parameter [9] as it only depends upon the physical parameters of the device such as oxide thickness, silicon film thickness, and the permittivity of the oxide and silicon.

\[ r = \frac{\epsilon_{si}t_{ox}}{\epsilon_{ox}t_{si}} \]  

(3.5)
Using the appropriate boundary conditions, the continuous analytic model [9] was used to derive the drain current expression in the linear region of operation shown in equation 3.6.

\[ I_{ds,\text{linear}} = 2\mu C_{ox} \frac{W}{L} \left( V_G - V_T - \frac{V_{DS}}{2} \right) V_{DS} \quad (3.6) \]

Similarly, the drain current expression in the subthreshold region of operation is defined by equation 3.7, where \( \Delta \phi \) is the work function of the top and bottom gate with respect to the intrinsic silicon channel.

\[ I_{ds,\text{subthreshold}} = \mu \frac{W}{L} t_{si} q n_i \phi_t e^{\frac{V_G - \Delta \phi}{\phi_t}} \left[ 1 - e^{-\frac{V_{DS}}{\phi_t}} \right] \quad (3.7) \]

A very important parameter known as the thermal voltage is denoted by equation 3.8 below. This quantity is the dominant temperature dependency for the MOSFET threshold voltage parameter. This term is present in the BJT and MOS (subthreshold) current equations.

\[ \phi_t = \frac{kT}{q} \quad (3.8) \]

For a \( V_{DS} \) greater than several thermal voltages, the subthreshold equation can be approximated by equation 3.9. This allows for simplification of calculations and improved linearity of the generated temperature components.

\[ I_{ds,\text{subthreshold}} = \mu \frac{W}{L} t_{si} q n_i \phi_t e^{\frac{V_G - \Delta \phi}{\phi_t}} \quad (3.9) \]
3.4.1 Comments on Drain Current

One way to improve the performance of the MOS device is to improve the carrier mobility or the saturation velocity of the channel [10]. Drain current positively correlates with carrier mobility, therefore increasing mobility improves the device current-voltage characteristics. Increasing the saturation velocity increases the applied gate electrode voltage range where the current linearly increases. Applying a voltage beyond this range is power-inefficient as the change in current is minimal.

Reducing the thickness of the gate oxide is extremely important as it reduces the short-channel effects (SCEs) and maximizes charge transport in the channel with decreasing power supply voltages [10]. As gate oxide thickness is decreased, there is an increased control over the channel through the gate electrode. The subthreshold slope parameter is useful as a measure of device efficiency due to the indirect control over the channel; it is not relevant at the drain and source electrodes. The electrical strength of the inversion layer is determined by three primary capacitances [10]: the depletion capacitance, the inversion capacitance, and the oxide capacitance. Thus, minimizing the oxide thickness increases the oxide capacitance which, in turn, yields an increase in inversion layer strength. A stronger inversion layer allows for the transport of more charge, as more charge is moved for a unit time and the drain current increases.

The most common way the double-gate device is used is to switch both
gates simultaneously (shorted gate configuration). However, biasing the back gate can be used to shift threshold voltage [10]. Adaptive regulation is possible by biasing the back gate. This biasing technique may be necessary in applications sensitive to process shifts. In symmetric double-gate devices with a channel thickness greater than 5 nm, the two inversion layers are independent of each other and there are two separate inversion charge peaks [10]. As the channel thickness is reduced, the two channels begin to merge into a single channel. Subthreshold slope is near ideal since the two gates provide better control over the channel, minimizing SCEs.

3.4.2 Comments on FinFET Subthreshold Operation

Subthreshold applications are gaining popularity in areas where low power is absolutely critical to the operation of the device such as in RFID tags, hearing aids, and pace-makers. A drawback of subthreshold operation is that delay is heavily dependent on power supply voltage [11]. Careful consideration is required to ensure the power supply is stable during switching transitions. Subthreshold operation of the FinFET device is a critical area of research as the push for low-power circuits often requires devices to be biased in the subthreshold region of operation. The main challenge lies in the speed of the device; subthreshold devices are significantly slower [3] than devices biased in strong inversion since there is less current available to charge and discharge parasitic capacitances within a circuit.
Currently, the majority of devices are built for operation in strong inversion [11]. To be used in high-frequency applications, subthreshold devices would need to be optimized. Retrograde and halo doping are used to suppress short-channel effects, specifically drain-induced barrier lowering and punch through, and to allow for threshold voltage adjustments. These implants may not be necessary as devices operating in subthreshold typically have low drain-to-source potential drops which in turn reduces the impact of drain-induced barrier lowering and body punch through. By removing the retrograde doping and halo doping profiles from the device, there is a reduced parasitic capacitance at the source and drain ends of the channel. FinFET fin thickness and gate oxide thicknesses can be optimized to improve speed in subthreshold operation as well as the switching energy of the device [3]. Gate capacitance is significantly smaller in subthreshold operation compared to operation in strong inversion [11]. This is due to the gate capacitance being a function of oxide capacitance and depletion capacitance whereas, in strong inversion, the gate capacitance is a strong function of oxide capacitance. Minimizing parasitic capacitances is crucial in subthreshold design since the low on-current, due to being in subthreshold conduction, reduces the maximum frequency the device can operate at for a given bias condition.

It has been proven that long-channel double-gate devices can be used in subthreshold operation without sacrificing performance [11]. In strong
inversion, delay is proportional to the square of gate length whereas in sub-threshold, the delay is a linear function of gate length. Since gate capacitance in a subthreshold double-gate MOSFET is nearly negligible, the delay is primarily a function of the on-current of the device [11]. Increasing the gate length does not result in a significant penalty on delay, which allows for larger gate lengths for better matching purposes. Overall, this leads to a more robust design. The on-current has a greater variation in short-channel devices compared to long-channel devices due to SCEs [11]. Therefore, it is wise to use long-channel devices in subthreshold operation. Double-gate MOSFETs have improved operating characteristics in subthreshold operation when compared to traditional planar CMOS devices since double-gate devices have little to no intrinsic capacitance in subthreshold operation. Double-gate devices have an extremely thin body between the two gates, which gives way to a higher degree of control over the channel and reduces the source and drain p-n junction capacitances [11].

Mismatch is reduced by increasing gate area; however, it was noted that increasing the number of fingers had a greater impact compared to increasing the number of fins. Decreasing the fin width improved SCEs; typically the fin width is smaller than the channel length. Mismatch and process variations were reduced due to the intrinsic channel which minimized random dopant fluctuations—one of the largest sources of mismatch error in planar CMOS devices. FinFET technology makes subthreshold operation easier to implement due to the double-gate structure which improved control over the
channel, reducing SCEs.

The analysis of the operation and modeling of the vertical double gate is critical as it justifies the need for a disruptive process solution. The numerous benefits offered by the vertical double-gate architecture outweigh the negative aspects associated with the technology shift. SCE suppression allows for the continuation according to Moore’s law towards a smaller minimum gate length while maintaining adequate performance. It is apparent that vertical double-gate devices, if properly designed, surpass the maximal performance margins of traditional bulk CMOS devices.

3.5 Process, Temperature, and Voltage Considerations

Monte Carlo analysis is the statistical variation of parameters that have an impact on the circuit. This is performed to ensure that yield will be high and that the device being fabricated will work as expected if a high number of runs meets yield requirements. A significant factor for the subthreshold bandgap in Monte Carlo simulations is threshold variation. Threshold variation on the diode-connected PMOS transistors has a significant effect since the devices operate in the weak inversion region, where the current equation shows a stronger dependence on threshold voltage. The ideal square law—which holds in strong inversion and saturation—has a squared dependence on threshold voltage, whereas in weak inversion, threshold variation has an exponential influence.
3.5.1 Comments on Process, Temperature, and Voltage Effects

The variations significantly affecting the performance of the device are threshold voltage variations, oxide thickness variations, and channel length variations [3]. In long-channel devices with a subthreshold swing close to the ideal value of approximately 60mV/decade at room temperature, current mismatch is primarily due to threshold voltage fluctuations [12]. This implies that the mismatch is independent of gate voltage fluctuations. In short-channel devices where subthreshold swing is a strong function of channel length, the threshold voltage fluctuations are not the only component in current mismatch [12]; it is also due to fluctuations in the subthreshold swing. In FinFET devices, the threshold voltage fluctuations are due to random dopant fluctuations (RDF). The mismatch is minimal in FinFET devices if properly designed, as the reduced density of dopants in the fin tends to mitigate the impact of RDF [12] which traditionally has been the most significant source of current mismatch. Other sources of mismatch must be considered due to the reduced magnitude of RDF [12]. Several other sources of mismatch would include line-edge roughness (LER) and metal-gate work function variations.

Typically, the drain current mismatch in the subthreshold region of operation is calculated by the standard deviation of threshold voltage. This would not be an accurate measure of drain current mismatch in short-channel FinFET devices since threshold voltage fluctuations are not the only significant
mismatch component, due to reduced presence of random dopant fluctuations [12]. However, in long-channel FinFET devices, it is a valid assumption that mismatch is a function of threshold voltage fluctuations as it is the dominant contribution.

The variation in gate length is one of the biggest contributors to process variations [6], more so than the silicon film thickness, which sets the channel width. Typically, to improve the performance of the drain current in a FinFET, a reduced gate length is required [12]. This is a design tradeoff as reduced gate length increases the prevalence of the SCEs. Increasing the gate length improved the subthreshold swing—the larger the gate, the closer to ideal value. The critical length is a function of the FinFET structure. The fin width has the most significant impact [12] on the critical length. Minimizing fin width is extremely important as it helps control the SCEs and it reduces the mismatch effects in the subthreshold region of operation. The critical length of the device [12] was defined to be at a 10% variation from the ideal subthreshold swing. Reducing critical length allows for a smaller minimum feature size.

In regions of operation where the channel has a significant degree of inversion, current variations are not as significant [12] compared to the average drain current. In the subthreshold region of operation, not only is there an exponential dependence, but the typical average current values are significantly lower than in the case of moderate or strong inversion. This implies that the mismatch of drain current between two FinFET devices can
be very significant, possibly up to one decade apart.

Traditionally, heavily doped polysilicon is used as the gate material [13]. With an intrinsic body, the threshold voltage for an NMOS device is negative and the threshold voltage for a PMOS device is positive. This results in depletion mode devices. Power consumption drastically increases since the devices cannot be turned off. This is solved in planar devices by doping the channel. While this can be done in FinFET devices, it is beneficial to change the work function of the gate rather than to implant the channel. The channel is extremely small and as a result of ion implantation, there is a pronounced increase in mismatch effects due to random dopant effects. An intrinsic channel reduces the impact of RDF. The challenge is finding a suitable metal gate to replace the traditional polysilicon gate. The intrinsic channel also gives way to higher mobility as the high channel doping concentrations degrade the mobility. Carrier mobility is higher in undoped channels; however, for planar CMOS devices, substrate control of the channel increases with increasing doping which reduces the effects of DIBL. This effect is negated by adding the back gate. Manipulation of threshold voltage by choosing the gate work function is not feasible in planar CMOS devices due to the significant increase in SCEs.

FinFET structures still abide by Pelgrom’s law, which states that the variance is inversely proportional to gate area. However, it was determined that the variation of the effective device width of several parallel minimum width devices is noticeably smaller compared to a single device of the same total
width [13]. Therefore, it is wise to increase the number of fingers on a device rather than to increase the number of fins on a device. Device width is quantized by the number of fins; more fins increases the channel width.

Threshold voltage significantly increases as body thickness decreases [13]. As body thickness decreases, the gates become closer together and there is an increase in control over the channel, which reduces SCEs and threshold voltage roll-off, which occurs when scaling down the length of the device. In symmetric double-gate devices with long channels and a thick body, there are two identical, independent inversion channels formed when the device is turned on [13]. A quantum well forms between the two gates once the body thickness significantly decreases. This introduces a potential barrier that needs to be surpassed, which intuitively explains the increase of threshold voltage.

Subthreshold swing has a strong dependence on body thickness [13]. As body thickness decreases, the threshold voltage rises and the channel becomes harder to invert. DIBL increases rapidly with increasing body thickness. It was shown that a 20% increase in gate oxide thickness resulted in an increase of subthreshold leakage of 13% [3]. Using an SOI device as a model, intuitively this makes sense as the term for subthreshold swing appears in the gate-to-source voltage term of the subthreshold drain current equation. The parameter arises as a direct consequence of indirectly controlling the surface potential of the device. As oxide thickness increases, the subthreshold swing increases due to a decreased control of the gate.
3.6 Planar CMOS Versus FinFET

As device scaling continues, SCEs become more prevalent, and measures must be taken to ensure proper device operation for sub-50 nanometer processes. One process solution emerging as a viable method to suppress SCEs is the FinFET device. A FinFET is a symmetric vertical double-gate device which is a drastic shift from the traditional planar CMOS devices. New equations were developed to accurately model the drain current in the subthreshold, linear, and saturated modes of operation. Sensitivity to statistical variation of device parameters across process, temperature, and voltage fluctuations is not as significant in FinFET processes when compared to bulk CMOS processes. Furthermore, vertical double-gate devices offer significant improvement in low-power applications, specifically within sub-threshold operation.

The push towards non-traditional architectures such as FinFETs and other multiple independent gate devices is due to the desire to continue scaling as predicted by Moore’s Law. Reducing the minimum gate length below 50 nanometers requires great care to minimize undesirable SCEs. Certain limitations that prevent planar CMOS technology from scaling below 50 nanometers are quantum tunneling through the thin gate oxide, quantum tunneling from source to drain and drain to body, as well as doping density issues to control threshold voltages [10]. The FinFET was carefully designed to overcome these limitations.
Chapter 4

Operational Amplifier Design

As the operational amplifier block within a bandgap is a crucial component, careful consideration was required in selecting the optimal topology. Several different amplifier and output stage topologies were examined for use in the bandgap and there were several key metrics that were considered when selecting the proper amplifier such as input common-mode range, gain, stability, and voltage headroom.

4.1 Amplifier Considerations

Figure 4.1: Amplifier Block Diagram
4.1.1 Output Stages

Figure 4.2 shows the typical implementation of a class A output stage. This implementation was used in the final design of the error amplifier. The benefits of this topology include simplicity and reduced voltage headroom requirements. The disadvantages of the class A output structure implemented were higher power dissipation and low gain; devices could not be cascaded due to voltage headroom limitations.

![Class A Output Diagram]

Figure 4.2: Class A Output

4.1.2 ICMR

The input common-mode range (ICMR) of the amplifier is an important parameter to keep in mind. If the common mode voltage is set near the limits of the ICMR, there is a high possibility that dropouts may occur if the transistors of the error amplifier fall out of saturation. A dropout is the condition where the reference voltage significantly falls below the average
reference voltage value. This occurred at low temperatures. At high temperatures, the higher order effects of the temperature coefficient become more prevalent and if not properly managed, increased quite rapidly—this was more pronounced with longer channels in the current mirror devices.

4.1.3 Effective Load Simulations

When designing the various blocks of the bandgap reference, it is important to simulate with an equivalent load. In the case of the error amplifier, the block was simulated with three PMOS capacitors and an NMOS device. The PMOS devices were sized the same as the current mirrors within the bandgap, and the NMOS device was sized the same as the start-up transistor to replicate the load. The devices $M_1$, $M_4$, $M_5$, and $M_{14}$ in Figure 4.3 correspond to the devices in Figure 5.17. In the ideal simulation, the bias node was set using an ideal voltage source. In the non-ideal simulations, the amplifier was simulated within the bandgap using stability (stb) analysis to break the feedback loop.

![Figure 4.3: Effective Load Simulation](image)
4.1.4 Gain

High DC gain for the operational amplifier is desirable for reducing the effects of power supply transients as well as to ensure that the output reference voltage of the bandgap design does not have significant temperature coefficient (TC) nonlinearities. The gain of the operational amplifier aids in reducing the variation of the reference voltage due to power supply variation, that is the gain of the amplifier improves the power supply rejection (PSR).

4.1.5 Stability

Stability is an important parameter to check using transient analysis. As there are many feedback loops, it is entirely possible the bandgap may become unstable if a large enough compensation capacitor is not present for multi-stage amplifiers, which are almost always required due to the low intrinsic gain of the aggressive sub-micron process [14].

4.2 Current Design Limitations

As minimum gate lengths continue to scale down there are two major design limitations, amplifier non-idealities and voltage headroom limitations. Amplifier non-idealities include finite gain, finite bandwidth, CLM, and output range limitations.

The impact of finite gain in the error amplifier can be derived using basic
circuit analysis. The following steps show the analysis of an error amplifier within a Brokaw cell bandgap. The schematic and derivation of the Brokaw cell can be found in Appendix C.

\[ V_{C1} - V_{C2} = \Delta V_C \]  \hspace{2cm} (4.1)

\[ \mu \Delta V_C = V_{BG} \]  \hspace{2cm} (4.2)

Rearranging this equation gives a more meaningful expression.

\[ \Delta V_C = \frac{V_{BG}}{\mu} \]  \hspace{2cm} (4.3)

The equation above shows the impact of finite gain in the error amplifier.
As the open loop gain decreases, the mismatch in voltage at the collectors of the BJTs increases. This gives way to an unequal current between the two branches. This mismatch in current manifests itself in the $\Delta V_{BE}$ component. Thus, the impact of finite gain is an increase or a decrease in the the $\Delta V_{BE}$ component.

The impact of finite unity gain bandwidth relates to stability. The bandgap reference could potentially become unstable due to finite unity gain bandwidth and injected noise. Even though the bandgap is a DC circuit, it will be connected to other circuits where AC capacitive coupling may be present, introducing noise on the bandgap voltage reference. Additionally, low bandwidth lengthens the start-up time for the circuit, increases settling times, and degrades PSR at high frequency.

CLM is undesirable in op-amp design because it degrades the intrinsic gain of the device. In strong inversion-saturation, the inverse of the slope is the resistance of the channel which is related to the Early Voltage. As CLM effects become more pronounced, the drain-to-source resistance decreases which in turn decreases the gain of the device. Additionally, CLM negatively impacts the matching of current mirrors. The effects of CLM can be reduced by increasing the physical gate length of the device.

If the output range of the error amplifier is not sufficient, it is possible that across certain corners, the amplifier will fail to provide the correct voltage to the PMOS current mirrors $M_1$, $M_2$, and $M_3$, which is shown in figure 5.2. This would result in drastic temperature variations. Therefore, it must
be ensured that the amplifier range falls within the range of gate voltages required for all corner and Monte Carlo variations.

Voltage headroom limitations due to sub 1-V power supplies eliminate many of the options available for the reduction of these non-idealities. Typically cascoding would be implemented in a bandgap design however; this is not feasible in sub 1-V power supplies.

4.3 Solutions to Limitations

There are several different solutions that are typically used in low-voltage, low-power designs. MOS-based diode rail-to-rail output stages can be implemented in processes with limited voltage headroom. Additionally, helper devices can be added to the op-amp to ensure that it remains in saturation at the worst case corners.

4.3.1 Offset

In the basic differential amplifier with an active load, the diode load introduces an imbalance of currents. Therefore, there will always be a systematic offset present. Offset can be reduced by using a fully differential amplifier with a common-mode feedback network (CMFB) to bias the connected gates of the first stage active load. Reducing input offset reduces the nonlinearities in the reference voltage temperature coefficient.
4.3.2 Robustness Techniques

Helper devices can be implemented to ensure devices remain in saturation when supply voltages are low. Figure 4.5 details one such method. $M_{3a}$ and $M_{4a}$ can be added to sink current, ensuring that the differential input remains in saturation across all corners.

![Figure 4.5: Differential Input Helper Transistors](image)

Figure 4.5: Differential Input Helper Transistors
4.3.3 Output Stages

The Monticelli Bias structure in figure 4.6 is a class AB output stage used for low power and rail-to-rail output applications [1]. This can be placed within a folded cascode, but it is wise to insert the structure in both branches of the folded cascode to ensure that the currents in the two branches are symmetric. This is important to minimize offset. $V_{in}$ provides a voltage input by the use of a common source device however; it is common to place the $M_3$ and $M_4$ diode in series within the branch of the cascode structure. This diode implementation is preferable in low-power applications as the voltage drop required is only $V_{ov}$. 

Figure 4.6: Monticelli Bias Schematic [1]
4.4 Implemented Error Amplifier Design

A simple amplifier, shown in figure 4.7, with a helper device was chosen. The helper device was necessary due to the supply voltage limitations. The implementation of the helper device, $M_8$, in the error amplifier was similar to the helper transistors, $M_{3a}$ and $M_{4a}$, in Figure 4.5. Both configurations attempt to keep the differential input in saturation. The basic differential amplifier topology was selected due to voltage headroom considerations.

The helper device is typically off, or sourcing little current. Once a dropout occurs, the gate voltage of the PMOS device decreases, which increases the overdrive of the helper device, sourcing more current, forcing the bandgap back into the proper state. The dropout issue was found to occur as a result of the differential input $V_{SD}$ of $M_1$ and $M_2$ falling below
the saturation voltage. The $V_{GS}$ of $M_3$ increased as temperature decreased, which pushes the input pair out of saturation. $M_8$ can also be seen as raising the $V_S$ of the input pair by clamping to $V_{DD}$ in order to increase the input pair $V_{SD}$. Before adding the device, over 50 Monte Carlo runs failed; afterwards every run passed at the expense of a 3 dB increase in PSR. The analysis of the helper device can be found in Appendix E.

Low threshold voltage devices were substituted for the PMOS differential pair to ensure the devices remain in saturation. Non-minimum length PMOS differential input devices were used because at low temperatures, the $V_{SD}$ dropped significantly, which causes the device to leave saturation resulting in poor output reference performance due to the degradation of output resistance. Wide devices for the PMOS source coupled pair ensured that the DC gain requirements were met across corners and Monte Carlo Analysis. Long devices were realized for the differential and output stage current sources due to the fact that long devices were required to keep the current sources in the saturation region. The tradeoff for increasing the length of the current source devices appeared at high temperatures, particularly past 100 °C. The reference voltage temperature coefficient significantly increased due to the error amplifier active load diode falling below saturation conditions due to increased headroom requirements.

For stability purposes, source-drain shorted capacitors were used, which have very little capacitance, but can be used in compensation networks,
start-up circuits, and as bypass capacitors in circuits such as current mirrors. The benefit of this is vast savings in area. These capacitors can easily be a factor of ten smaller than MIM capacitors.
Chapter 5

Bandgap Design

The design of a bandgap reference requires careful consideration of several parameters. The primary objective is to create an output reference voltage that ideally does not change with temperature. In addition to a temperature-insensitive output, there are other considerations as well, such as die area, power dissipation, device mismatch, and ease of trimming [15].

![Figure 5.1: Temperature Dependency Superposition](image)

The bandgap reference sums a proportional-to-absolute-temperature (PTAT) and complimentary-to-absolute-temperature (CTAT) temperature component together to create a temperature-insensitive reference, which is a zero-to-absolute-temperature reference (ZTAT).
Figure 5.2: Implemented 14 nm Bandgap Reference

Figure 5.2 shows the final implementation of the bandgap designed in the 14 nanometer FinFET technology for this thesis. By summing a PTAT component and a CTAT component together, a temperature-insensitive output current was mirrored through the $M_1$-$M_3$ PMOS current mirror network. The $M_3$ current creates a potential drop across $R_4$, generating the output reference voltage of the bandgap. $R_1$ is used to set the quiescent current of the bandgap and generate the PTAT dependency and $R_2$ and $R_3$ are used to adjust the temperature coefficient of the output reference.
5.1 Simulation to Calculate $\frac{\partial V_{SG}}{\partial T}$

The temperature dependence, $\frac{\partial V_{SG}}{\partial T}$, varies from process to process. It is important to simulate the temperature coefficient for use within the theoretical calculations to arrive at the optimal resistor ratios in the bandgap circuit. Figure 5.3 shows the typical method used to simulate the diode TC. The current flowing through the diode should be approximately what is expected for the diode in the bandgap implementation as the TC does have a dependence on the DC operating point.

![Figure 5.3: MOS Diode Voltage Temperature Coefficient Simulation](image)

5.2 Headroom Issues in Cascoded Current Mirrors

Using the ideal square law for planar CMOS devices, an expression for the gate bias voltage in a cascoded current mirror (Figure 5.4) was derived to gain a better understanding of the impact of $V_{OV}$ and the aspect ratio.
Figure 5.4: NMOS Cascode Current Mirror

\[ V_{GS3} = \sqrt{\frac{2I_{D3}}{k'(\frac{W}{L})_3}} + V_T \] (5.1)

\[ I_{D1} = \frac{1}{2} k' \left( \frac{W}{L} \right)_1 \left( V_{G1} - V_{GS3} - V_T \right)^2 \] (5.2)

\[ I_{D1} = \frac{1}{2} k' \left( \frac{W}{L} \right)_1 \left( V_{G1} - \sqrt{\frac{2I_{D3}}{k'(\frac{W}{L})_3}} - 2V_T \right)^2 \] (5.3)

\[ \sqrt{\frac{2I_{D1}}{k'(\frac{W}{L})_1}} = V_{G1} - 2V_T - \sqrt{\frac{2I_{D3}}{k'(\frac{W}{L})_3}} \] (5.4)

\[ V_{G1} = 2V_T + \sqrt{\frac{2I_{D1}}{k'(\frac{W}{L})_1}} + \sqrt{\frac{2I_{D3}}{k'(\frac{W}{L})_3}} \] (5.5)
Assuming equal currents and aspect ratios for M1 and M3,

\[ V_{G1} = 2V_T + 2\sqrt{\frac{2I_{ref}}{k'(W/L)}} \] (5.6)

Taking overdrive voltage into consideration, a smaller aspect ratio is preferred because the device will be biased well into saturation, where minor changes in overdrive voltage do not have as significant of an impact on the output of the current mirror. Overdrive, \( V_{ov} \), is the difference between the applied gate-to-source voltage and the threshold voltage.

\[ V_{OV} = V_{GS} - V_T \] (5.7)

Therefore, to reduce the \( V_{G1} \) needed to bias the current mirror devices in saturation, the aspect ratios of \( M_1 \) and \( M_3 \) need to be relatively large to minimize the overdrive voltage. However, a large overdrive voltage is desired as it reduces the impact of small gate voltage fluctuations on the drain current. The design tradeoff is variability for voltage headroom.

Increasing the output impedance through increasing the lengths of the devices reduces the effects of CLM and also provides benefits for the physical design with regards to matching. Furthermore, the increased impedance of the cascode structure provides benefits to the bandgap PSR since the output resistance is significantly improved. Equation 5.8 shows the approximate resistance of the cascode structure.
\[ r_{out} \approx r_o (1 + g_m r_o) \quad (5.8) \]

Unfortunately, in sub 1-volt supplies, cascoding is not usually a viable method to employ, which can be seen from 5.6. The minimum required gate voltage occurs when the aspect ratio approaches infinity. This is still two threshold voltages, which consumes a significant portion of the voltage headroom. For this reason, cascoded current mirrors were not used in this work.

### 5.3 Sizing Simple Current Mirrors

The simple current mirror was used in the bandgap reference implemented because minimum voltage headroom requirements were difficult to meet with a cascoded structure. Equation 5.9 below governs the necessary aspect ratio of the devices for a simple current mirror, assuming that the reference current and the overdrive voltage desired are known. This is a straightforward calculation for a first pass at device sizing.

\[ \left( \frac{W}{L} \right) = \frac{2I_D}{k'V_{ov}^2} \quad (5.9) \]

It is important to consider \( V_G \) because the voltage cannot go below the negative supply rail. The current mirror must be sized such that the steady state operating point is satisfied by the error amplifier output voltage range for all corners and Monte Carlo cases.
5.4 Temperature Dependency

The temperature coefficient is defined by equation 5.10 below; it is a measure of the stability of the voltage reference with respect to temperature fluctuations.

\[ TC = \frac{\Delta V_{ref}}{\Delta T \times V_{ref,nom}} \times 10^6 \text{ppm/°C} \]  

(5.10)

All devices have temperature effects associated with them. The most commonly used devices are resistors and transistors.

5.4.1 Temperature Dependence of Resistors

It is typical that metal resistors have a significant temperature dependence and a very low resistance. \(P^+\) polysilicon doped resistors tend to have the lowest dependence on temperature and have moderate values of
sheet resistance. Due to these desirable characteristics of $P^+$ doped polysilicon, it is commonly used in a low temperature coefficient design, such as in a bandgap reference. Well resistors have a higher temperature dependence when compared to polysilicon resistors. However, well resistors usually have smaller process variations compared to polysilicon resistors due to higher density.

5.4.2 Temperature Dependence of Field Effect Transistors

There are several different physical parameters of the MOSFET that have an impact on the operation of the device. These parameters are oxide thickness, effective channel length, effective channel width, threshold voltage, mobility, and substrate doping. All of these parameters also have a direct impact on the temperature dependence of the MOSFET device.

5.4.3 PTAT Generation

The difference between gate-to-source voltages of two MOS diodes with different current densities is PTAT and exhibits an approximately linear relationship with respect to temperature. This derivation of the differential $V_{GS}$ relationship is shown in Appendix B. This relationship is identical to the BJT diode relationship.
5.4.4 CTAT Generation

The CTAT component is generated from the diode. Two different diodes can be created for a bandgap reference. The traditional implementation uses a BJT. The base and collector are shorted together and with a high value of current gain, an insignificant amount of current flows through the base [15]. The second method that can be used to create a diode is to short the gate and drain of a MOSFET. However, in a bandgap reference where subthreshold operation is required to reduce the nonlinearities in the CTAT coefficient, there is a significant dependence on current flowing through the device on the gate-to-source voltage—this leads to increased variation in reference voltage.

5.5 Device Sizing Considerations

In an effort to reduce mismatch, many devices were placed in parallel and self-cascoded. Current mirror devices were sized to ensure that the devices remained in saturation for all corners and Monte Carlo cases. It was noted that undersizing the current mirror significantly increased the number of dropouts due to the fact that devices required more voltage to remain in saturation. The current flowing in the output branch of the circuit approached zero as the MOSFET approached the cut-off region. MOSFETs with large gate lengths exhibited more gate leakage, which adversely impacted bandgap performance.
5.6 Drain Voltage Equalization Current Mirror Bandgap

In the standard implementation of a drain voltage equalization current mirror bandgap, a diode is used in series with $R_2$ to reduce power dissipation, reduce the effects of CLM, and decrease the passive resistor area;
however, this reduces the voltage headroom of the bandgap.

Referring to the beta-multiplier bandgap in Figure 5.6, the derivation of output reference voltage and quiescent branch current are straightforward.

\[ V_m = V_{SG4} \]  

(5.11) \[ V_p = V_{SG5} + IR_1 \]  

(5.12) 

With a high enough error amplifier gain, the positive and negative nodes of the amplifier are equal.

\[ V_{SG4} = V_{SG5} + IR_1 \]  

(5.13) \[ V_{SG4} - V_{SG5} = IR_1 \]  

(5.14) \[ \Delta V_{SG} = IR_1 \]  

(5.15) 

Provided that the MOSFETs are running in weak inversion, as shown in Appendix B.

\[ \Delta V_{SG} = \phi t \ln(N) = IR_1 \]  

(5.16) \[ I_{ref} = \frac{\phi t \ln(N)}{R_1} \]  

(5.17) \[ V_{ref} = I_{ref} R_2 + V_{SG6} \]  

(5.18) \[ V_{ref} = \frac{R_2}{R_1} \phi t \ln(N) + V_{SG6} \]  

(5.19) 

Furthermore, differentiating the reference voltage and equating the slope
to zero should be done to determine the resistor ratio required to ensure the output reference has no dependence on temperature.

\[
\frac{\partial V_{\text{ref}}}{\partial T} = \frac{R_2}{R_1} \ln(N) \frac{\partial \phi_t}{\partial T} + \frac{\partial V_{SG6}}{\partial T} = 0
\] (5.20)

\[
\frac{R_2}{R_1} \ln(N) \frac{\partial \phi_t}{\partial T} = -\frac{\partial V_{SG6}}{\partial T}
\] (5.21)

\[
\frac{R_2}{R_1} = -\frac{\partial V_{SG6}}{\partial T \ln(N) \frac{\partial \phi_t}{\partial T}}
\] (5.22)

\(\frac{\partial \phi_t}{\partial T}\) is a known quantity, \(\frac{\partial V_{SG6}}{\partial T}\) can be experimentally determined, \(R_1\) is found by assuming a current to meet power specifications, and \(R_2\) is found by using the ratio derived above.

The resistor in series with the larger diode can be adjusted to control the current flow in the circuit. Larger values of resistance reduce the quiescent current while lower values of resistance increase the quiescent current of this circuit due to Ohm’s law.

The error amplifier ensures that the drain voltages of the current mirror are identical and that the current flowing in either branch is also identical. Finite error amplifier gain will result in inaccuracies in mirrored drain current due to CLM effects. The \(\Delta V_{SG}\) component may not be adequately forced across \(R_1\) which would result in non-linearities in the TC.

It is important to ensure that the error amplifier used has low input-referred offset and a high PSRR as these two terms have a significant impact on the error of the output reference voltage. The operational amplifier can
be biased using a constant transconductance bootstrapped bias network to ensure line variations do not significantly impact the loop-gain of the amplifier. Alternatively, using a PMOS differential pair, the PMOS current source can be biased off of the current mirrors within the bandgap reference. Input-referred offset is typically attributed to the differential input pair as the offset is gained up throughout the signal path of the circuit. The most common reasons as to why input offset is non-zero is due to CLM and mismatched input devices. In other words, the current flowing in the two branches of the symmetric differential pair is not exactly equal.

Minimum supply voltage imposes limitations on the minimum voltage headroom needed by the current mirrors to operate in strong inversion-saturation. If the current mirrors are not operating in strong inversion-saturation, there will be errors in the mirrored output current. Typically, the drain to source voltage of the current mirrors must be greater than the overdrive voltage of the device to remain in saturation. The drain-to-source voltage of the PMOS diodes must be greater than several thermal voltages to ensure the device is saturated in subthreshold conduction for the subthreshold drain current approximation to hold.

Another bootstrap biasing technique that can be used, which has a low TC and low variability with respect to the power supply, is a resistive divider inverter-clamp network connected to the PMOS CM gates. The resistive divider will bias the current mirror of the error amplifier. The main two drawbacks to this topology are increased silicon area and slow response
time, which is why it was not implemented in the final design.

The addition of a unity gain buffer on the output of the bandgap can alleviate the effects of loading the bandgap. The TC and input offset of the output buffer will introduce further error into the output reference voltage.

There are three main considerations when sizing devices within a bandgap voltage reference. Sizing devices with the objective to minimize power and area, optimization of device sizes to minimize mismatch and process variations, and sizing the diodes and resistors to achieve a zero-to-absolute-temperature (ZTAT) TC.

$R_1$ sets the current for each of the branches. The sizes of the current mirror devices do not set the current, if properly sized. Therefore, the aspect ratios of the current mirror devices should be made small enough that the current mirror is in strong inversion-saturation for the worse case corner or Monte Carlo run. Similarly, the PMOS diodes do not dictate the current if sized properly. The larger the device is, the further into subthreshold the diodes will operate. It is important to appropriately size the diode because the second diode is $N$ times larger. If the devices are extremely wide, the gate-to-source voltage across them will decrease, but the current will still be set by the resistor.

It is important to increase output impedance of the current mirrors because it alleviates the effects of CLM. Additionally, it also increases the PSR of the voltage reference. Relating to the traditional cascoded current mirror, the aspect ratio of the bottom NMOS diode should be made large to
minimize the overdrive voltage. However, a large overdrive voltage is often desirable as it reduces the impact of small gate voltage fluctuations on the drain current. The reason as to why a small overdrive voltage is desirable in this low power design is due to the fact that it reduces the minimum output voltage to keep all devices in saturation and functioning properly as current mirrors.

Figure 5.7: Bootstrapped Widlar Current Source

$M_5$ and $M_6$ function as MOS bypass capacitors and $R_1$ sets the bias current in the Widlar Current source. The top PMOS current mirror provides the reference current for the NMOS current mirror, and the NMOS current mirror provides the reference current for the PMOS current mirror. This is
known as a bootstrapped current source. The bootstrapped Widlar current source needs an external start-up mechanism to work properly. The circuit is metastable. It will eventually start up due to leakage currents however, it will take much too long to reach the correct operating state. In a bootstrapped circuit, the correct diode connection is in the path where the loop gain is less than unity, as both paths have positive loop gains and will only be stable if the loop gain is less than unity.

![Diagram](image)

**Figure 5.8: Inverter Based Start-Up**

The first option for a start up is the inverter based start up as seen in Figure 5.8, the circuit injects current when the bandgap is in the incorrect operating state, and the start up shuts off once the correct state has been reached. The inverter trip point must be sized such that the circuit adequately turns off for the steady state $V_G$.

The second start-up option, seen in Figure 5.9 makes use of SCEs specifically—leakage currents. The leakage currents generated by the diodes will accelerate the start-up. A faster start-up time will require larger devices due to
reduced equivalent resistance to AC ground, which impacts the RC time constant. Each node in the current reference needs to be charged up. The drawback to this method is that static power consumption can increase significantly in low power systems.

Figure 5.9: Leakage Start-Up Method

Figure 5.10: Dynamic Start-Up Circuit

Option three is a dynamic clamp start-up circuit as shown in Figure 5.10. The PMOS capacitor, $M_2$, prevents current flow through $M_1$. Initially, the
charge in the capacitor is zero according the capacitor-voltage relationship. Therefore, the gate of $M_3$ is at the positive supply voltage and the two nodes in the current bias network are clamped by $M_3$. Current begins to flow and the second equilibrium point is reached. At that point, $M_1$ is on and the gate of $M_3$ is at zero, effectively shutting off the start-up when it is no longer needed. This dynamic start-up does not dissipate a significant amount of static power. The capacitance should be made smaller, or the two NMOS devices larger to decrease the start-up time.

5.7 Sensitivity Issues

Within the drain voltage equalization current mirror bandgap, $\frac{R_2}{R_1}$ is a ratio, which tends to be more stable with process and mismatch variation compared to individual device variations. $\frac{R_2}{R_1} \phi_t \ln(N)$ does not vary much across process or mismatch, whereas $V_{SG6}$ in figure 5.6 will change significantly across process and mismatch variation due to subthreshold diode sensitivities. Therefore, the percent variation with respect to the mean can be reduced by increasing the average value of the first term in the reference voltage equation given by equation 5.19.

The beta multiplier bandgap reference is a good starting point for a design; however, the diode load on the output stage increases the minimum voltage needed for proper operation. This design would not be feasible with a sub-one-volt supply without modifications.
5.8 Modified FinFET Drain Voltage Equalization CM Bandgap

Initially, when the bandgap reference was simulated, extreme variation of the output reference voltage was experienced. This was due to voltage headroom issues. Looking at the $V_{DD}$ vs $V_{REF}$ curve, it was apparent that for the nominal supply voltage, it was operating in the region of the curve where the slope was steepest. In other words, the minimum voltage requirements were not met to ensure all devices were operating properly.
Dropouts occurred as a result of the insufficient voltage overhead. It is important to reduce the minimum operating supply voltage as much as possible if the design is to be compatible with sub-one-volt supplies. With large overdrive voltages, the voltage variations are minimized but the voltage overhead increases. By increasing the aspect ratio of select devices, their overdrives can be reduced and as a result, the minimum operating supply voltage will be reduced.

In the modified Bohannon reference [2], $R_2$ and $R_3$ are used to tune the temperature coefficient, $R_1$ is used to set the reference current, and $R_4$ aids in setting the reference voltage. Derivations for the resistor ratios can be seen in Appendix A.

Minimum operating voltage of the bandgap is very important in low power systems where the nominal supply voltage is often less than one volt. At low supply voltage, the op amp will have low gain, and there will be errors with $\Delta V_{SG}$ and the TC. Looking at the equations derived, it shows that the overdrive voltage of the current mirror has a significant impact on the minimum operating voltage of the circuit. Additionally, the diode also has a significant impact on this operating point, which was why low threshold devices were considered in this work. Appendix B proves that MOS diodes in weak inversion operate within a bandgap similarly to traditional BJT diodes. The BJT would not work because the minimum base-to-emitter voltage of the BJT, for a given current, is much larger than that of the minimum required gate-to-source voltage of the MOSFET. The gate-to-source voltage
of the PMOS diode can be reduced, assuming a fixed aspect ratio. The other handle on the drain current equation is the drain current. By reducing the drain current, in order to satisfy the equation, the gate-to-source voltage must also decrease. By reducing the drain current, the minimum operating voltage will be reduced as well, which is beneficial for voltage headroom and power, but the impact of noise and variation is more apparent.

The diode in the active load of the error amplifier is important to take into consideration as well. The overdrive voltage of the diode must be low, to reduce the minimum operating supply voltage where the amplifier functions properly. Increasing the gain of the error amplifier flattens out and reduces the total variation of the $V_{DD}$ versus $V_{ref}$ curves. The vertical shift of the curve for different power supply voltages is related to the PSRR of the circuit. The current mirror can be self-cascoded to increase the output resistance, which will improve PSR.

CLM is present in this design because the drains of $M_1$ and $M_2$ are at a different voltage compared to the drain of $M_3$. The bandgap was intentionally designed this way. Voltage headroom was determined to be more important than the error introduced into the TC. By removing the output diode, voltage headroom significantly improved and the majority of dropouts were fixed.

By implementing a MOS bandgap using subthreshold diodes, a low TC can be obtained while maintaining low power and reduced area. The cost associated with this is increased variability of the reference voltage due to
the subthreshold diodes. However, this can be mitigated by using resistive trimming networks, weighted current mirror trimming networks, and diode trimming networks.

The purpose of Monte Carlo analysis is to find the worst-case variation from the nominal reference voltage. This is important for determining if the system meets specifications and if it does not, it will aid in determining the number of bits required for trimming. Likewise, it is also used to find the worst case TC that can be trimmed equal to or lower than a certain specified TC.

The complete schematic including the error amplifier, start-up network, and bandgap can be seen in Figure 5.17.

5.9 DC Biasing Considerations

Two different bias mechanisms were considered for the amplifier, one biased off of the current mirrors in the bandgap, and one biased independently of the bandgap bias network. The bias network of the operational amplifier was designed to provide the necessary gate voltages to the current sources to ensure proper operation of the amplifier across corners and Monte Carlo analysis. A threshold reference was considered for the bias network. The NMOS device used to create the threshold voltage bias had a large aspect ratio to minimize the overdrive voltage, ensuring the bias voltage was placed
as close to the threshold voltage of the device as possible. The principle behind the threshold reference is that the output current is set by the ratio of the threshold of an NMOS transistor and the resistor at the gate of that NMOS device. With this, a fairly stable output current can be realized. Ultimately, the amplifier was biased off of the bandgap due to improved performance (no dropouts, improved PSR, and reduced area).

5.10 Resistive Trimming Networks

It is typical to perform Monte Carlo analysis to determine the total variation of the bandgap reference voltage in order to decide if trimming the output voltage is necessary. If the output variation exceeds the tolerable limit, methods must be used to shift the reference voltage to the appropriate value.

A trimming network was considered due to the fact that planar CMOS and FinFET devices have significant threshold variations in subthreshold conduction. One way to mitigate this was to introduce combinations of parallel resistors to trim the output to a specific voltage by tying the gates high or low. With regards to the resistor trimming network, the output resistive network does not have a significant impact on the temperature coefficient. The resistors need to be optimized to reduce area and to be able to trim a wide range of reference voltages. Six resistors were used in this bandgap trim network; however, that may not be ideal due to area considerations. In
theory this circuit will be able to eliminate that variation, converging to a minimum value.

Binary-weighted PMOS devices could be used as a different trim method. MOS switches were used to switch the devices on or off. The results were similar to the resistive trim; however, the advantage of this method was a drastic reduction in area and an increased voltage headroom.

When using a MOSFET as a switch in a trimming network, it is important to make the devices large to ensure that the drain-to-source voltage drop is negligible during proper circuit operation.

Figure 5.12: Series Trim Network Schematic
Resistive trimming networks are designed to adjust the output reference voltage to the correct reference value. Diode trimming networks could also be employed. A diode trimming network focuses on reducing the overall temperature coefficient. Mismatch between the two diodes introduces TC
error in the output reference voltage, and this can be corrected with careful trimming.

The series trim network of Figure 5.12 only allows for shifting the reference up, the parallel trim network of Figure 5.13 only allows for shifting the reference down, and the effective network of Figure 5.14 provides the capability of shifting in both directions.

5.11 Applications

![Figure 5.15: Precision Current Reference](image)

One application of the bandgap reference would be the precision current source. The output current in Figure 5.15 is a function of the load resistor, device aspect ratio, and the bandgap voltage. With an external resistor, or a carefully trimmed resistor, the reference current will be very close to the desired value and would have temperature-insensitive properties. This
could be extended to a binary-weighted trimmed network current source to allow for tuning of the output current for high precision applications. The reference could also be used within an analog-to-digital converter (ADC) or a phase-locked loop (PLL).

5.12 Matching

Mismatch also has a significant effect on the output reference voltage. Mismatch and process variations in the middle current branch of the bandgap cause the most significant variation in current, due to the fact that the branch sets the output current and thus the output reference voltage. Mismatch between the two diodes causes a difference in $V_{SG}$ required to equalize the current differential. The voltages at the differential input stage of the operational amplifier vary with this change in diode voltage due to mismatch, which in the case of an increased differential voltage will cause an increased output voltage in the operational amplifier, reducing the output current and thus introducing positive variation into the reference voltage. With a decreasing differential voltage due to diode mismatch, the output reference voltage has a negative variation from the mean.

Active resistors take up significantly less area compared to passive resistors; however, they are more susceptible to process variations, especially
in low-power applications where subthreshold operation is typical. Furthermore, active based resistors and capacitors contain non-linear characteristics. In precision applications, it is wise to design using passive devices.

With regards to resistor matching, it is important to only use one type of resistor and a unit length for that resistor [15]. Additionally, it is wise to make use of resistor ratios rather than absolute values of resistance, as ratios are much more resistant to process variation when compared to individual values of resistance. This is due to the proximity effect [15]. That is, the process gradients tend not to change very much in a local area. Furthermore, dummy resistors can be used around the edges of the device to mitigate edge effects. The nodes of the dummy devices can be tied to ground or the power supply rail. This method of reducing mismatch can also be applied to capacitors, BJTs, and MOSFETs. Guard rings can also be employed to reduce the impact of substrate noise on the circuit. Similar to the dummy devices, it is best not to leave the nodes floating. Nodes can be tied to ground or the power supply rail.

Connecting several resistors in parallel decreases the mismatch associated with the equivalent resistance; four parallel unit resistors can reduce mismatch by half [15]. This can only be done to a certain extent before larger sets of parallel combinations do not yield high returns due to the presence of systematic offset. Eventually spatial gradients will dominate when random mismatch is reduced [15]. Methods to reduce spatial gradients must
then be put to use. These methods include common centroid layout and interdigititation. Several parallel and series combinations can be used to reduce the systematic offset; the correct placement of the devices in relation to one another will reduce the impact of spatial gradients.

5.13 Higher Order Curvature Correction

The parabolic temperature dependence can be attributed to the higher order temperature dependencies embedded within mobility. The proposed compensation network minimizes the variation of these higher order temperature dependencies.

![Figure 5.16: Higher Order Curvature Correction Schematic](image)

PSR is greatly reduced due to the near unity gain inversion stage in the compensation network. PSR of the compensation network in figure 5.16 can be increased by removing the diode connected device $M_2$ by biasing it
through a constant transconductance block. Equation 5.23 below shows the potential drawback of this: it may result in a non-unity gain inversion as a result of mismatched transconductances. Aside from poor PSR, the other design tradeoff is increased power dissipation. The higher order correction circuit inverts the concavity of the input reference. The input reference and the inverted input reference are halved and then summed together. With the principle of superposition, the halved inverted reference added to the halved input reference ideally results in the reference voltage with no dependence on temperature. Appendix D shows the derivation of the resistor ratios of the output block.

\[ A_v = -G_{in}R_{out} = -g_{m1}R_{o2} \approx -\frac{g_{m1}}{g_{m2}} \]  \hspace{1cm} (5.23)
5.14 Generalized Bandgap Design Methodology

1. Theoretical Analysis of the Bandgap
   - Derive output reference voltage
   - Take the derivative of the output reference voltage, equate to zero, and solve for important ratios

2. Simulation for Process Parameters
   - Simulate for parameters such as $\frac{\partial V_{SG}}{\partial T}$

3. Solve for Resistor Ratios
   - Optimize for power and area
   - Optimize for process and mismatch variation

4. Simulate $V_{DD}$ vs. $V_{ref}$ Graph
   - Ensure that there are no voltage headroom limitations on the bandgap

5. Sweep Temperature to Verify the Output Reference

6. Histogram Plots of all Important Voltages Over Corner Extremes and Monte Carlo
   - Current Mirror $V_{SG}$
   - Differential Input $V_{SD}$
   - Amplifier Current Mirror
• Amplifier Active Load

• Output Reference Voltage

7. Transient Analysis

• Ensure start-up

• Inject noise to ensure stability

Figure 5.17: Complete Bandgap Schematic
Chapter 6

Bandgap Reference Results

Cadence tools were used to simulate and verify theoretical analysis of both the FinFET bandgap at the 14-nanometer node, and the planar CMOS bandgap at the 45-nanometer node. The same bandgap was implemented and optimized for each process. This was done to allow for a comparison between traditional CMOS and FinFET technologies.

6.1 14-nm Drain Voltage Equalization CM Bandgap

The FinFET-based bandgap was simulated using several different methods in order to properly verify the functionality and robustness of the circuit. PVT analysis, Monte Carlo analysis, DC sweep analysis, transient analysis, and AC analysis were all used in the verification of the design.

6.1.1 PVT Results

After successful completion of the nominal simulations, the PVT simulations were performed as a way to ensure that the bandgap would function
as expected when subjected to process variations during the fabrication of the device. Figure 6.1 shows the corresponding reference voltages for the process extremes. Process variations were selected as $nn$, $ss$, $ff$, $sf$, and $fs$. The first letter is reserved for the NMOS corner and the second is reserved for the PMOS corner. The $n$ corner is nominal, $s$ is slow, and $f$ is fast. Aside from $nn$, the other four corners are what the manufacturer has deemed to be the $3\sigma$ limits of expected process variations. Power supply variations were set as $\pm$ 10% and the temperature range was set as -20°C to 125°C.

![Figure 6.1: 14 nm Reference Voltage PVT Variations](image)

The majority of the PVT traces operated as expected. They were relatively flat excluding higher order parabolic temperature dependencies. However, several corners began increasing at higher temperatures at the slow NMOS corner, raising the TC of the output reference for that corner. This
was attributed to voltage headroom issues at high temperatures. The gate-to-source voltage of the NMOS active load diode decreased with increasing temperature to the point where the active load performance degraded. The gain degradation of the error amplifier also introduced TC nonlinearities in the output reference. This introduced input-referred offset within the error amplifier resulting in a PTAT error in the output reference voltage. This problem was exacerbated as channel lengths were increased; therefore, it was imperative to find an optimal gate length with respect to voltage headroom and device mismatch. If channel lengths are increased without maintaining the aspect ratio, the drain current decreases and with less current flowing through the active load diode, the $V_{GS}$ drops quicker, increasing the PTAT error introduced into the output reference.

6.1.2 Monte Carlo Results

Monte Carlo analysis was performed to assess the impact of process variation and device mismatch. PVT analysis does not account for device mismatch and only simulates the process extremes. There were three corner cases that were used within the Monte Carlo analysis which were the three power supply corners ($0.9V_{DD}$, $V_{DD}$, and $1.1V_{DD}$). For each of the three corners of the analysis 100 Monte Carlo runs were simulated.

Figure 6.2 shows that the temperature coefficient was relatively insensitive to process variation. The output reference exhibits curvature uniformity. This was expected since the TC was set by resistor ratios, not an individual
value of resistance, which was more stable when subjected to variations.

![Figure 6.2: 14 nm Reference Voltage Monte Carlo Analysis](image)

Initially, there were many dropouts when running Monte Carlo simulations. The bandgap current mirror must be sized appropriately, that is, the channel length must be made longer and possibly self-cascoded. The error amplifier must be sized properly, the channel length must be increased for the differential input pair, the second stage bias mechanism, and the diode load of the first stage to minimize voltage headroom requirements, ensuring all devices remain in saturation.

The dropouts occur in the bandgap current mirror when the drain-to-source voltage falls below the minimum voltage required for saturation. The current mirror no longer works properly—resistance drops and current is not mirrored properly. This causes issues with the TC as well as the DC operating point of the reference voltage. For the amplifier, the input pair, diode,
and current source drop in resistance, so the gain falls drastically which will in turn alter the reference voltage. It was noted that by increasing the channel lengths of the current mirror, the number of dropouts reduced and variation of the reference voltage was reduced, which aligned with expected results.

From Figure 6.2 it can be seen that the bandgap is relatively insensitive to temperature changes, but the average reference voltage varies significantly. Bandgaps are temperature independent, but they are not independent of process, voltage, and mismatch effects. Therefore, it is imperative to reduce these effects as much as possible.

6.1.3 $V_{DD}$ vs. $V_{ref}$

The output reference graph with respect to power supply is an important simulation that should be used to verify that the bandgap has sufficient voltage headroom. Figure 6.3 shows the required supply voltage for proper operation of the bandgap reference with the resistive trim network. Figure 6.4 shows a similar graph; however, it is for the MOS-based trim network. This analysis should be done for both the 45 worst case corners as well as the 300 run Monte Carlo analysis to ensure all runs have sufficient headroom.

From Figure 6.3, it can be seen that the minimum supply voltage required to satisfy the voltage headroom requirements is approximately 600 mV. Once voltage headroom requirements were met, the reference voltage did not significantly change with respect to further increases in the power
supply voltage. This implies that the magnitude of the bandgap PSR is relatively high.

Figure 6.3: 14 nm Voltage Headroom with the Resistive Trim Network

Figure 6.4 shows that with the MOS trim network, it requires less voltage headroom as a result of removing the series-connected resistive trim network. The minimum supply voltage required to satisfy the voltage headroom requirements in the MOS trim configuration is approximately 525 mV. Unlike the resistive trim network, several corners significantly increased at high supply voltages. The MOS trim network creates several parallel MOS devices from the power supply to the reference voltage which reduces the impedance of the output branch, thus lowering the magnitude of the bandgap PSR. Ultimately, the resistive trim network was chosen for the final design of the bandgap due to the improved PSR. The design tradeoff was increased
area due to the passive resistors, and increased voltage headroom requirements.

![Figure 6.4: 14 nm Voltage Headroom with the MOS Trim Network](image)

**6.1.4 PSR Graph**

AC analysis was used to generate the PSR graph. By applying a small amplitude ac sinusoidal source to the power supply, the PSR was measured as gain from the power supply to the output reference voltage. In this bandgap implementation, the main areas in which PSR was dictated is the PMOS current mirrors and the error amplifier. Cascoding the PMOS current mirrors would significantly improve the PSR; however, this could not be done due to voltage headroom limitations. Similarly, improving the error amplifier gain would benefit the overall bandgap PSR. Furthermore, pre-regulation could also be used to improve PSR but this would introduce a voltage drop on the supply which would negatively impact the issue of
voltage headroom. A simple RC filter was used as a way to increase the magnitude of the PSR; however, the drop across the resistor resulted in reference voltage dropouts within Monte Carlo simulations. Even though this method provided a 3 dB increase in PSR, it was not implemented in the final bandgap design due to the introduced voltage headroom limitations.

Figure 6.5: 14 nm PSR Graph

Figure 6.5 shows the 14 nm bandgap PSR response. At DC the bandgap attenuates signals from the path of the power supply to the output reference voltage. As frequency increases, the magnitude of attenuation decreases until the peak of approximately unity PSR. The DC gain of the error amplifier should be improved in order to improve the bandgap PSR at low frequencies since cascoding cannot be used. Additionally, improving the gain-bandwidth product of the amplifier will improve the high frequency PSR. It is critical to increase the magnitude of attenuation at high-frequencies as
most dynamic circuits such as PLLs or clock networks inject noise with frequency content near the carrier frequency. It is desirable to have an adequate PSR at high-frequencies, since the majority of noise injected on the bandgap power supply will consist of high-frequency content.

Increasing the lengths of the PMOS current mirrors will shift the second pole location to a lower frequency resulting in an improvement in the worst case PSR since the first pole cancels out the zero and the second pole introduces a negative 20 dB/decade slope.

6.1.5 Transient Analysis

Transient analysis to verify start-up is typically done to ensure that the bandgap reference start-up circuit turns on and settles within a specified margin. It also ensures that the bandgap is stable. The DC analysis provides the correct DC solution even if the bandgap is not stable. Transient analysis is one of several ways to ensure that the circuit will always be stable when subjected to process variation and device mismatch.

Figure 6.6 shows the start-up analysis of the 14 nm bandgap. The power supply was ramped from 0 V to $V_{DD}$ and an initial condition was placed on the bandgap to ensure that it was off at the beginning of the ramp. The corners that were most likely to fail the start-up analysis were the cold temperature, slow corners. All traces properly started due to the leakage start-up method implemented in the bandgap. It can also be seen that the bandgap is stable as all traces settled and did not oscillate. A stability analysis was
successfully conducted using the *stb* analysis tool to verify stability.

Figure 6.6: 14 nm Transient Start-Up Analysis

Figure 6.7 shows the digital transient switching of the 64 different combinations of resistors within the trim network. There is a slight loss of uniformity in step size that results from the series-parallel combination of resistors; however, this tradeoff was acceptable as it allowed for trimming to increase and decrease the reference voltage.
6.1.6 Trim Results

There were two different trim networks that were designed and simulated for the bandgap: the six-bit resistive trim network and the six-bit MOS trim network. The performance of the resistive trim network was more accurate compared to the MOS trim network at the expense of area and voltage headroom. Ultimately, the resistive trim network was selected to ensure the TC of the bandgap was minimized.
Figures 6.8 and 6.9 show the results of trimming a single reference voltage. In both cases, the TC does not significantly change with the 64 trim codes. However, it can be seen that the resistive trim network had a significantly lower temperature coefficient when compared to the MOS trim network. The trim range of the resistive network was 204.1 mV, which was triple the trim range of the MOS network. The resistive trim network provided higher accuracy in terms of temperature coefficient as well as the safety factor of having a larger trim range which would be useful if the reference voltage was significantly different from what was expected after fabrication.
6.2 Higher Order Compensation

The equation for drain current in a subthreshold MOSFET diode has a dependence on mobility. There are higher order temperature dependencies embedded within mobility that are not canceled out by the linear PTAT and CTAT components. The best case TC that was obtained for the 14 nm bandgap was $66.46 \frac{PPM}{^{\circ}C}$, for applications that require more precise references, higher order compensation is an absolute necessity.
Figure 6.10: 14 nm Compensation Comparison

Figure 6.10 shows the comparison of the results of the uncompensated and compensated output reference voltages. The proposed compensation network passed 300 runs of Monte Carlo and was found to be relatively insensitive to process and mismatch variations. The design tradeoff associated with this higher order compensation network is reduced attenuation of low-frequency power supply noise (PSR) and increased power dissipation. Another benefit of this compensation network is that it does not reduce the operating temperature range of the bandgap. Second-order piecewise compensation networks shorten the operating temperature range because the output reference voltage begins to rapidly increase at the temperature extremes as a result of the addition of the piecewise compensation network.

Leakage through the MOS capacitor and the second stage MOS gate was
significant enough to introduce offset in the output amplifier of the compensation network. The compensation network required low offset for accuracy. This issue was fixed by attaching a dummy MOS capacitor and a dummy second stage to balance the gate leakages in the two branches of the differential stage. An isolation buffer was used at the output of the bandgap to ensure that the compensation network did not introduce an unwanted temperature coefficient by drawing current. Another isolation buffer was used at the output of the concavity inversion block which served two purposes, to eliminate current draw from $M_2$ and to provide a symmetric input load to both $R_1$ resistors.

The results shown in Figure 6.10 imply that it is important to invest effort in higher order compensation for a high-performance voltage reference.

### 6.3 45 nm Drain Voltage Equalization CM Bandgap

The design of the 45 nm bandgap reference was completed to allow for comparison of a similar design in FinFET technology (14 nm) and planar CMOS technology (45 nm). The comparison of the bandgaps show that the better design was the 14 nm node. Even with a lower power supply voltage and smaller minimum gate length, the design process was easier at the 14 nm node and the performance exceeded the bandgap at the 45 nm node. This was attributed to the numerous benefits of the FinFET architecture that were previously stated.
PVT Results

Comparing the PVT results of the two bandgaps, the FinFET bandgap output reference voltage did not vary as much from trace to trace. The results of the 45 nm bandgap PVT simulation in figure 6.11 show that the TC of the 45 nm reference varies more significantly than the 14 nm bandgap.

Figure 6.11: 45 nm Reference Voltage PVT Variations

6.3.1 Monte Carlo Results

The analysis shown in Figure 6.12 for the 45 nm bandgap once again shows increased variation of trace-to-trace output reference voltage. The TC also varies significantly across Monte Carlo. The TC has higher order nonlinear components that are not present in the 14 nm bandgap. The 14 nm bandgap trace appears to be parabolic in nature whereas the trace of the 45 nm bandgap inverts concavity approximately halfway through the
temperature sweep.

Figure 6.12: 45 nm Reference Voltage Monte Carlo Analysis

6.3.2 \( V_{DD} \) vs. \( V_{ref} \)

The 45 nm process is a 1 V process, the design was simulated at 0.9 V, 1 V, and 1.1 V. The minimum voltage headroom required is 900 mV which can be seen in the \( V_{DD} \) vs. \( V_{ref} \) plot shown by Figure 6.13. The FinFET architecture had an additional 100 mV of voltage headroom which provided extra headroom for the process extremes. Once the supply voltage reached the minimum value required for the output reference to flatten off, the reference did not significantly increase with further increases in power supply voltage. The slope of the reference past \( V_{DD,\text{min}} \) was greater in the 45 nm bandgap compared to the 14 nm bandgap. This implies that the PSR of the 14 nm bandgap outperforms that of the 45 nm bandgap.
6.3.3 PSR Graph

The PSR of the 45 nm bandgap was significantly worse when compared to the 14 nm bandgap. Figure 6.14 shows the PSR response. The DC PSR
value is almost 20 dB higher than the 14 nm bandgap and the worst case PSR value is approximately 10 dB higher than the worse case of the 14 nm bandgap. The only redeeming factor of the PSR plot at 45 nm node is the fact that the worst-case PSR occurs at a higher frequency.

### 6.3.4 Transient Analysis

The transient analysis of the 45 nm reference shown in Figure 6.15 verifies that all traces start up as expected and the output reference is stable. The transient start-up analysis was conducted similarly to the 14 nm bandgap. A power supply was ramped up from 0 V to $V_{DD}$ with an initial condition on the bandgap to ensure it was initially off.

![Figure 6.15: 45 nm Transient Start-Up Analysis](image-url)
6.3.5 Trim Results

The results of the resistive trim network analysis are shown in Figure 6.16. In the case of the 45 nm bandgap, the output TC does have a slight dependence on the various combinations of the resistive trim network. This was attributed to the fact that the voltage headroom is 100 mV lower than in the 14 nm bandgap. Additionally, the traces are not as evenly spaced out when compared to those of the 14 nm bandgap resistive trim network.

![Figure 6.16: 45 nm 6-Bit Resistive Trim Network](image)

6.4 Layout Comparison

Figure 6.17 shows the comparison of the BJT and MOS diode layouts. Figure 6.17.a.a shows the equivalent nine BJT diodes required for a typical BJT-based bandgap implementation and 6.17.a.b shows the finalized implementation of the two MOS diodes in the FinFET bandgap. It is apparent
that the MOS diode greatly reduces the required area of a bandgap. Furthermore, the BJT devices shown in the figure are minimum sized devices. To properly conduct current larger BJTs must be used. One such implementation of a BJT-based bandgap [2] used 81 BJT devices for the diodes. As power supply voltages continue to decrease to accommodate decreasing minimum gate lengths more BJT diodes will be required to conduct current. At a certain point it will no longer be feasible to design using BJT diodes. It is imperative to invest effort into MOS diode-based implementations of bandgaps. Figure 6.17.b.a shows the drastic increase in BJT area while 6.17.b.b shows that the area of the MOS device did not change.

Figure 6.17: Layout Comparisons
### 6.5 Comparative Bandgap References

<table>
<thead>
<tr>
<th>Specifications</th>
<th>[Current Work (14 nm)]</th>
<th>[Current Work (45 nm)]</th>
<th>[2]</th>
<th>[16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Node [nm]</td>
<td>14</td>
<td>45</td>
<td>65</td>
<td>130</td>
</tr>
<tr>
<td>$V_{DD,nom}$ [V]</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>$V_{DD,min}$ [V]</td>
<td>0.6</td>
<td>0.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$V_{ref,min}$ [mV]</td>
<td>370.5</td>
<td>507.2</td>
<td>650</td>
<td>290.3</td>
</tr>
<tr>
<td>$V_{ref,avg}$ [mV]</td>
<td>412.2</td>
<td>572.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$V_{ref,max}$ [mV]</td>
<td>458.7</td>
<td>630.5</td>
<td>669.6</td>
<td>416.4</td>
</tr>
<tr>
<td>$\sigma_{V_{ref}}$ [mV]</td>
<td>17.1</td>
<td>21.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$V_{ref,TrimRange}$ [mV]</td>
<td>204.1</td>
<td>85</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$TC_{min}$ [PPM $^{\circ}$C]</td>
<td>66.46</td>
<td>65.21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$TC_{avg}$ [PPM $^{\circ}$C]</td>
<td>153.6</td>
<td>138</td>
<td>251.0</td>
<td>240</td>
</tr>
<tr>
<td>$TC_{max}$ [PPM $^{\circ}$C]</td>
<td>230.2</td>
<td>267.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_{TC}$ [PPM $^{\circ}$C]</td>
<td>19.19</td>
<td>39.33</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$TC_{comp}$ [PPM $^{\circ}$C]</td>
<td>6.25</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$P_{dissipation,avg}$ [$\mu$W]</td>
<td>48.8</td>
<td>50.6</td>
<td>37.0</td>
<td>0.9</td>
</tr>
<tr>
<td>$t_{ramp,min}$ [$\mu$s]</td>
<td>30</td>
<td>50</td>
<td>10,000</td>
<td>50</td>
</tr>
<tr>
<td>$A_v$ [dB]</td>
<td>43.5</td>
<td>-</td>
<td>-</td>
<td>89.6</td>
</tr>
<tr>
<td>$PSR_{60Hz}$ [dB]</td>
<td>-42</td>
<td>-22.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Area [$\mu$m x $\mu$m]</td>
<td>-</td>
<td>-</td>
<td>202.2x198.1</td>
<td>90x60</td>
</tr>
<tr>
<td>Temperature Range [C]</td>
<td>-20 to 125</td>
<td>-20 to 125</td>
<td>-40 to 125</td>
<td>0 to 100</td>
</tr>
</tbody>
</table>

Table 6.1: Bandgap Results and Comparison

Comparing the four bandgaps in Table 6.1, the current works and [16] use subthreshold MOSFET devices; however, the current work at the Fin-FET node is at a significantly smaller technology node and power supply
voltage. Due to licensing issues, layout could not be performed in the FinFET technology. It can be seen that even though [16] is at a larger technology node, the total area is less than that of [2]. The estimated area of the current work at the FinFET node was approximately half that of the other MOS bandgap [16]. Of all four references, the current work of the FinFET bandgap had the lowest average temperature coefficient, for both the compensated and uncompensated reference.

The Bohannon version of the Banba bandgap had a TC of 251.0 ppm/°C and 2.1% total variation. The Bohannon bandgap works properly once the supply voltage is greater than 0.9 volts. The temperature range that was considered for the design was -40°C to 125°C. The 14 nm design had an average TC of 153.6 ppm/°C and had an operational temperature range of -20°C to 125°C. The minimum operating voltage was 300 mV lower than the Bohannon bandgap. These results clearly show that the implementation using MOS diodes is vastly superior to the BJT implementation as supply voltages continue to scale.

The power supply voltage of the 14 nm reference was 200 mV lower than the 45 nm reference and the FinFET bandgap was still able to outperform the planar CMOS bandgap. While the 45 nm bandgap had a lower average TC when compared to the 14 nm bandgap, the standard deviation was nearly double. This implies that the TC after fabrication of the 45 nm reference would not be as well controlled as the 14 nm reference. This was also the case with output reference voltage. The 45 nm reference had nearly
double the standard deviation compared to the 14 nm reference. Additionally, another reason as to why the average TC of the 45 nm reference is lower than the 14 nm reference is due to the fact that the reference voltage is approximately 100 mV greater. The change in voltage of a single trace was comparable in both bandgaps however, with a higher average reference value, the TC appears to be lower.
Chapter 7

Conclusions

The primary focus of this thesis was to create a bandgap using subthreshold FinFET diodes at the 14 nm node to address voltage headroom limitations. The same design was also created at the 45 nm node to allow for a direct comparison between planar CMOS and vertical double gate technologies.

A simple-two stage amplifier was chosen for the error amplifier in an effort to minimize the required voltage headroom. The main drawback of this topology was limited gain since cascoding could not be used due to the low supply voltage.

The MOSFET bandgap operated properly with a minimum supply voltage of 0.6 V for the resistive trim configuration and 0.5 V for the MOS-based trim configuration. While there was significant variation in the output reference due to process variations, the use of a trimming network was justified by the drastic reduction in area. The minimum temperature coefficient was $66.46 \frac{PPM}{^\circ C}$. The average temperature coefficient of the higher order correction method was $6.25 \frac{PPM}{^\circ C}$. Comparing the various bandgap designs,
it was clear that overall, the design in the FinFET process was superior to the BJT based bandgap and the planar CMOS bandgap.

Overall, this thesis shows that it may be necessary to design bandgaps using techniques, such as subthreshold diodes, to reduce the required voltage headroom and a FinFET process may be needed to achieve the required performance specifications.

7.1 Transition From Planar CMOS to FinFET

The transition from planar CMOS to FinFET was relatively easy. The benefits offered by the FinFET structure allowed for the implementation of a high-performance bandgap reference. The device variation of the design at the 45 nm node was more significant than the device variation of the FinFET bandgap. Table 6.1 shows the standard deviations associated with the TC and the output reference. In both cases the 45 nm bandgap exhibited a wider spread. This seems to imply that designs in a FinFET technology may have higher yield due to the lower variance.

The main reasons as to why it was easier to design in the FinFET process were reduced CLM and higher intrinsic gain. Reduced CLM allowed for better matching of the PMOS current mirrors in the bandgap. Mismatched current mirrors introduced TC non-linearities. Higher intrinsic gain implies that a two-stage error amplifier has a higher gain in the FinFET process compared to the 45 nm CMOS process. This led to improvements in the
TC by ensuring that the drain voltages of the PMOS current mirrors were equivalent. The improved error amplifier gain also led to improvements in the PSR of the bandgap. Reduced CLM is associated with intrinsic gain in the sense that as CLM effects decrease, the output resistance of a MOSFET in saturation increases which, in turn, improves the gain of the device. This increase in output resistance also aids in improving the bandgap PSR.

7.2 Future Work

Future work should be performed to reduce bandgap variation by investing effort in process-insensitive diode feedback circuits. The higher order compensation network would benefit from pre-regulation to improve PSR.

The error amplifier should be redesigned to include a rail-to-rail constant transconductance input to improve operation since the common-mode voltage puts the differential input close to subthreshold operation. Additionally, a fully differential cascoded topology should be investigated. A fully differential cascode stage would eliminate any diode-connected devices which greatly increase the minimum voltage required for operation. The cascoded amplifier would provide higher gain which would improve the PSR of the bandgap.

Lastly, an all-MOS bandgap with no resistors should be explored as that would allow for reduced voltage headroom, especially at extremely low currents. The $V_{ov}$ required for saturation in subthreshold conduction is only
several $\phi_t$. An all-MOS bandgap would have a significant reduction in area since passive resistors require a significant amount of space.

7.3 Implications

This thesis shows that FinFET devices can and should be used at aggressive submicron technology nodes for optimal circuit performance. Furthermore, as technology continues to scale down, it is important to continue exploring other methods that can be used to reduce the required voltage headroom of a circuit. This thesis shows that it is possible to replace BJT diodes with MOS diodes without degradation of circuit performance. MOS diodes in subthreshold operation offer two benefits: reduced power dissipation and reduced voltage headroom. The drawback of the MOS diode is increased sensitivity to process variation.
Appendix A

Derivation of Resistor Ratios for the DVECM Bandgap

The derivation of the resistor ratios is exactly the same as the derivation of the resistor ratios in the Bohannon bandgap because the MOS diodes operate the same as the BJT diodes.

Figure A.1: Implemented 14 nm Bandgap Reference
\[ I_{ds, subthreshold} \approx I_0 \frac{W}{L} e^{\frac{V_{SG}}{\phi_t}} \]  
(A.1)

\[ I_{R1} = \frac{(V_{SG1} - V_{SG2})}{R_1} \]  
(A.2)

\[ \Delta V_{SG} = \phi_t \ln(N) \]  
(A.3)

\[ I_{R1} = \frac{\phi_t \ln(N)}{R_1} \]  
(A.4)

\[ I_2 = I_{R1} + I_{R3} \]  
(A.5)

\[ I_{R3} = \frac{(V_{SG1} - V_{ref})}{R_3} \]  
(A.6)

\[ R_2 = R_3; I_{R2} = I_{R3}; I_1 = I_2 = I_3 \]  
(A.7)

\[ I_{R4} = I_{R1} + I_{R3} + 2I_{R3} = I_{R1} + 3I_{R3} \]  
(A.8)

\[ I_{R4} = I_{R1} + 3I_{R3} = \frac{\phi_t \ln(N)}{R_1} + 3\frac{(V_{SG1} - V_{ref})}{R_3} \]  
(A.9)
\[ V_{ref} = I_{R4}R_4 = R_4 \left[ \frac{\phi t \ln(N)}{R_1} + 3 \left( V_{SG1} - V_{ref} \right) \right] \]  
(A.10)

\[ V_{ref} = \frac{\phi t \ln(N)R_2R_4 + 3V_{SG1}R_1R_4}{R_1R_2 + 3R_1R_4} \]  
(A.11)

\[ R_4 = KR_1; R_2 = MR_1 \]  
(A.12)

\[ V_{ref} = \frac{MK\phi t \ln(N) + 3KV_{SG1}}{M + 3K} \]  
(A.13)

\[ \frac{\partial V_{ref}}{\partial T} = \frac{K}{3K + M} \left[ M \frac{\partial \phi t \ln(N)}{\partial T} + 3K \frac{\partial V_{SG1}}{\partial T} \right] = 0 \]  
(A.14)

\[ M \frac{\partial \phi t \ln(N)}{\partial T} = -3K \frac{\partial V_{SG1}}{\partial T} \]  
(A.15)

\[ M = -3K \left[ \frac{\partial V_{SG1}}{\partial \phi t \ln(N)} \right] \]  
(A.16)

This equation theoretically provides a ZTAT output reference for the bandgap. Starting with (A.13), the resistor ratio K can be solved for.

\[ K = \frac{MV_{ref}}{3V_{SG1} + M\phi t \ln(N) - 3V_{ref}} \]  
(A.17)
If $V_{SG1} = V_{ref}$,

$$K = \frac{V_{SG1}}{\phi_l \ln(N)}$$  \hspace{1cm} (A.18)

However, if $V_{ref}$ is greater than $V_T$, the diodes will not be in weak inversion and there will be significant non-linear temperature dependencies. Therefore, current must flow through $R_2$ and $R_3$ if $V_{ref}$ is greater than $V_T$. 
Appendix B

Derivation of $\Delta V_{GS}$

\[ I_{ds,\text{subthreshold}} = \mu \frac{W}{L} t_{si} q n_i \phi_t e^{\frac{V_{GS} - \Delta \phi}{\phi_t}} \left[ 1 - e^{\frac{-V_{ds}}{\phi_t}} \right] \]  
(B.1)

\[ I_{ds,\text{subthreshold}} \approx \mu \frac{W}{L} t_{si} q n_i \phi_t e^{\frac{V_{GS} - \Delta \phi}{\phi_t}} \]  
(B.2)

The FinFET is a symmetric vertical double gate device which implies that both gates have the same work function therefore, $\Delta \phi = 0$.

\[ I_{ds,\text{subthreshold}} \approx \mu \frac{W}{L} t_{si} q n_i \phi_t e^{\frac{V_{GS}}{\phi_t}} \]  
(B.3)

\[ I_{d1} = I_{d2} \]  
(B.4)

\[ \left( \frac{W}{L} \right)_1 e^{\frac{V_{GS1}}{\phi_t}} = \left( \frac{W}{L} \right)_2 e^{\frac{V_{GS2}}{\phi_t}} \]  
(B.5)

\[ \frac{e^{\frac{V_{GS1}}{\phi_t}}}{e^{\frac{V_{GS2}}{\phi_t}}} = e^{\frac{V_{GS1} - V_{GS2}}{\phi_t}} = \left( \frac{W}{L} \right)_2 \left( \frac{W}{L} \right)_1 \]  
(B.6)
\[
\frac{\left(\frac{W}{L}\right)_2}{\left(\frac{W}{L}\right)_1} = N \quad (B.7)
\]

\[
\ln \left[ e^{\frac{v_{GS1} - v_{GS2}}{\phi_t}} \right] = N \quad (B.8)
\]

\[
\frac{\Delta V_{GS}}{\phi_t} = \ln(N) \quad (B.9)
\]

\[
\Delta V_{GS} = \phi_t \ln(N) \quad (B.10)
\]

This derivation proves that the MOS diode operating in weak inversion results in the same equation as the BJT diode for the differential current density component of the bandgap.
Appendix C

A Review of the Brokaw Cell

Figure C.1: Brokaw Cell Schematic
The derivation of the bandgap voltage is relatively straightforward and is shown below.

\[
I_{C1} = A_1 I_s e^{qV_{BE1}/kT} \tag{C.1}
\]

\[
I_{C2} = A_2 I_s e^{qV_{BE2}/kT} \tag{C.2}
\]

\[
\frac{I_{C1} A_2}{A_1 I_{C2}} = e^{q(V_{BE1} - V_{BE2})/kT} \tag{C.3}
\]

\[
\ln \left[ \frac{I_{C1} A_2}{A_1 I_{C2}} \right] = \frac{q(V_{BE1} - V_{BE2})}{kT} \tag{C.4}
\]

\[
\Delta V_{BE} = \frac{kT}{q} \ln \left[ \frac{I_{C1} A_2}{A_1 I_{C2}} \right] \tag{C.5}
\]

Then for equal currents,

\[
\Delta V_{BE} = \frac{kT}{q} \ln \left[ \frac{A_2}{A_1} \right] \tag{C.6}
\]

Using a ratio of eight, this can be simplified to:

\[
\Delta V_{BE} = \frac{kT}{q} \ln(8) \tag{C.7}
\]

\[
V_{E1} = 2I_{C1}R_2 \tag{C.8}
\]

Modifying the BJT current equation,

\[
\frac{I_{C1}}{I_s} = e^{q(V_{BG} - 2IC1R_2)/kT} \tag{C.9}
\]
Solving for the output voltage yields,

\[ V_{BG} = \phi_t \ln \left( \frac{I_{C1}}{I_s} \right) + 2I_{C1}R_2 \]  \hspace{1cm} (C.10)

\[ I_{C1} = \frac{\Delta V_{BE}}{R_1} \]  \hspace{1cm} (C.11)

\[ V_{BG} = \phi_t \ln \left( \frac{I_s e^{V_{BE}/\phi_t}}{I_s} \right) + 2 \frac{R_2}{R_1} \Delta V_{BE} \]  \hspace{1cm} (C.12)

The final expression can be written as:

\[ V_{REF} = V_{BG} = V_{BE} + 2 \frac{R_2}{R_1} \phi_t \ln(8) \]  \hspace{1cm} (C.13)

The area ratio between the two devices was set at eight, which is typical. This was done because it allows for a better layout design. The resistor value of \( R_1 \) was selected using equation C.14 below. Notice the significance of this equation: the current is set by this equation.

\[ R_1 = \frac{\phi_t \ln(8)}{I_{C1}} \]  \hspace{1cm} (C.14)

Differentiating the expression for the reference voltage, setting it to zero, and solving for \( R_2 \) yields the equation below.

\[ R_2 = \frac{-R_1 T}{\phi_t} \frac{\partial V_{BE}}{\partial T} \frac{1}{2 \ln(8)} \]  \hspace{1cm} (C.15)
Appendix D

Higher Order Compensation Output Block Analysis

Figure D.1: Higher Order Curvature Correction Schematic

\[ V_m = V_{Ref-Comp} \frac{R_3}{R_2 + R_3} \]  \hspace{1cm} (D.1)

\[ \frac{V_1 - V_m}{R_1} + \frac{V_2 - V_m}{R_1} = 0 \]  \hspace{1cm} (D.2)
\[ V_1 + V_2 = 2V_m \]  \hspace{1cm} (D.3)

\[ \frac{V_1 + V_2}{2} = V_{\text{Ref-Comp}} \frac{R_3}{R_2 + R_3} \]  \hspace{1cm} (D.4)

\[ V_{\text{Ref-Comp}} = \left[ \frac{R_2 + R_3}{R_3} \right] \frac{V_1 + V_2}{2} \]  \hspace{1cm} (D.5)
Appendix E

Analysis of the Error Amplifier Helper Device

Figure E.1: Two Stage Op-Amp Schematic With Helper Device $M_8$

\[ I_{d3} \approx \mu \frac{W}{L} t_s \eta_q \phi_t e V_{GS} = I_0 e^{V_{GS}} \]  \hspace{1cm} (E.1)
\[ V_{GS3} = \phi_t \ln \left( \frac{I_{D1}}{I_0} \right) \] (E.2)

\[ I_{D1} = I_{D3} = \frac{I_{D5} + I_{D8}}{2} \] (E.3)

\[ V_{GS3} = \phi_t \ln \left( \frac{I_{D5} + I_{D8}}{2I_0} \right) \] (E.4)

Therefore, when a dropout occurs due to the \( V_{GS} \) of the active load diode dropping to zero at cold temperatures, \( I_{D8} \) injects current through \( M_1 \) and into the diode \( M_3 \). Equation E.4 shows that the injected current would raise the active load diode \( V_{GS} \), pushing it back into the proper operating state.

In the event that the \( V_{SD} \) of the input pair falls significantly, \( I_{D8} \) turns on and begins clamping \( V_s \) towards the rail, raising \( V_{SD} \) by increasing the source voltage of the input pair.

When the bandgap is in the correct state, the helper device is in weak to moderate inversion where the off-current does not have a significant contribution to the amplifier.
Bibliography


