UNIVERSITY OF CALIFORNIA, SAN DIEGO

BiCMOS Sampling Circuits for High-Speed Data Converters

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Electrical Engineering (Electronic Circuits & Systems)

by

Kristian Norgaard Madsen

Committee in Charge:

Professor James F. Buckwalter, Chair
Professor Peter M. Asbeck
Professor Lawrence E. Larson

2013
The Thesis of Kristian Norgaard Madsen is approved and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2013
I’d like to dedicate this work to my parents, Lars and Pam Madsen. Without them, this would have been impossible.
Because what is your night worth without a story to tell? And why wield a word
like worth if you’ve got nothing to sell? People drop pennies down a wishing well, as
if the cost of a desire is equal to that of a thought. But if you’ve got expectations,
eventually others have bought your exact same dream for the price of a ‘hard work, hang
in, hold on’ mentality. Like, I accept any challenge so challenge me. Like, I brought
a knife to this gun fight, but the other night I mugged a mountain so bring it, I’ve
had practice.

Shane Koyczan - “The Crickets Have Arthritis”
# TABLE OF CONTENTS

Signature Page .............................................. iii
Dedication .................................................. iv
Epigraph ..................................................... v
Table of Contents .......................................... vi
List of Figures ............................................... viii
List of Tables ............................................... xii
Acknowledgements .......................................... xiii
Vita ............................................................ xv
Abstract ..................................................... xvi

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

1 Heterojunction Bipolar Transistors ................. 2
   1.1 The Bipolar Junction Transistor .................... 2
   1.2 The Heterojunction .................................. 6
   1.3 Heterojunction Bipolar Transistors ............... 8
   1.4 SiGe BiCMOS ........................................ 10
   1.5 InP HBTs and an InP/Si BiCMOS Technology ....... 13
   1.6 A Technology Comparison ......................... 17

2 Sampling Circuits and Systems ....................... 20
   2.1 Sampling Theory .................................... 20
   2.2 Sampling Circuits .................................. 21
      2.2.1 Track-and-Hold ................................. 22
      2.2.2 Sample-and-Hold .............................. 23
   2.3 Time Interleaved Sampling ......................... 25
      2.3.1 4-Phase Time Interleaved Track-and-Hold .... 26
      2.3.2 Signal Reconstruction ......................... 28
      2.3.3 Gain Mismatch .................................. 30
      2.3.4 Voltage Offset .................................. 31
      2.3.5 Systematic Timing Errors ...................... 33

3 Track-and-Hold Design .................................. 36
   3.1 Switched Emitter Follower Architecture .......... 36
   3.2 Emitter Follower Linearity Analysis ............... 38
   3.3 Droop Induced Non-Linearity ....................... 48
   3.4 Pedestal Error ...................................... 49
   3.5 Hold-Mode Feedthrough .............................. 50
   3.6 Sampling Jitter ..................................... 51

4 A 40GS/s High Resolution Track-and-Hold .......... 54
   4.1 High Resolution Compensation Techniques .......... 55
Figure 1.1: Current components in an NPN Bipolar Junction Transistor (BJT). ......................................................... 3
Figure 1.2: Three types of heterojunctions formed in semiconductors [2]. . 6
Figure 1.3: A Type II heterojunction where $E_{GA} > E_{GB}$. .................. 7
Figure 1.4: An abrupt Type II heterojunction. .................................. 8
Figure 1.5: Current components in an HBT after [6]. ............................ 9
Figure 1.6: Representative bandgap diagram of standard Si BJT (solid) and SiGe HBT with graded base (dashed). The diagram includes a plot of the germanium concentration in the base region, after [7]. 11
Figure 1.7: Simulated $f_t$ and $f_{max}$ curves from a range of devices sizes included in the IBM 9HP library. .......................... 13
Figure 1.8: A cross sectional diagram of the InP/Si BiCMOS stack up, reproduced from [13]. ........................................... 15
Figure 1.9: Image of an InP DHBT from the G4 library. © 2010 HRL Laboratories, LLC. All Rights Reserved. Reproduced with permission from [14]. .......................................................... 15
Figure 1.10: Simulated $f_t$ and $f_{max}$ curves for some of the InP devices in the BiCMOS library. ........................................... 16
Figure 1.11: Comparison of $f_t$ and $f_{max}$ versus emitter current density, for the InP and SiGe BiCMOS processes. ...................... 17
Figure 1.12: Device speed comparisons for SiGe HBTs over the past 10 years. 18
Figure 1.13: Device speed comparisons for InP HBTs over the past 10 years. 19
Figure 2.1: Discrete sampling of a continuous signal. (a) The continuous signal. (b) The sampling instants shown overlapping the continuous signal. (c) The remaining sampled data. ......................... 21
Figure 2.2: Operation of the track-and-hold, representing the continuous input signal, clock phases and output signal. ......................... 22
Figure 2.3: Two methods of producing a sample-and-hold. (a) Formed by cascading two track-and-holds. (b) Formed by two parallel track-and-holds fed into an analog multiplexer. ............................ 23
Figure 2.4: Operation of the sample-and-hold. The output of the single track-and-hold is shown in dashed lines. ............................. 24
Figure 2.5: Operation of the sample-and-hold modeled with an impulse train and rectangular impulse response. ............................. 24
Figure 2.6: Block diagram of a time interleaved sampling system with associated phase relationships. ............................................. 25
Figure 2.7: A 4-phase time interleaved track-and-hold proposed as an example for the following section. ......................................... 26
Figure 2.8: Time domain plot of the 4 overlapping track-and-hold phases, along with the output spectrum of a single track-and-hold. . 27
Figure 2.9: A 4-phase time interleaved track-and-hold with a multiplexed output. The effective sampling rate increases to $4F_s$. 28
Figure 2.10: Time domain plot of the reconstructed signal, along with the spectrum of a single track-and-hold and reconstructed signal. 29
Figure 2.11: Model of a time interleaved system where gain mismatch is present. 30
Figure 2.12: Output spectrum of a time interleaved sampling system where gain mismatch is present between channels. 31
Figure 2.13: Model of a time interleaved system where offset is present between channels. 32
Figure 2.14: Output spectrum of a time interleaved sampling system where offset is present between channels. 32
Figure 2.15: Model of a time interleaved system where systematic timing error is present between channels. 33
Figure 2.16: Output error incurred by clock timing errors, shown for 3 input tones. 34

Figure 3.1: The switched emitter follower architecture. 37
Figure 3.2: Operation of a switched emitter follower track-and-hold. (a) SEF in track-mode. (b) SEF in hold-mode. 37
Figure 3.3: Emitter follower driving a capacitive load, representing the SEF during track mode. 38
Figure 3.4: Non-linear small signal model of the capacitively loaded emitter follower. 38
Figure 3.5: Small signal block diagram of the feedback system. 40
Figure 3.6: Generation of the first-order operator. 40
Figure 3.7: Second-order loop model. 41
Figure 3.8: Generation of the second-order operator. 42
Figure 3.9: Models representing $H_{3,1}$. (a) Third-order loop model incorporating $G_3$. (b) Third-order term generated by the third order nonlinearity. 43
Figure 3.10: Models representing $H_{3,2}$. (a) Third-order loop model incorporating $G_2$. (b) Third-order term generated by the second order nonlinearity. 44
Figure 3.11: A comparison between simulated and theoretical distortion in the SEF. 48
Figure 3.12: Common-mode and differential droop. 49
Figure 3.13: The input signal leaks through the base-emitter capacitance of the switching device. 51
Figure 3.14: Representation of the amplitude error induced by timing uncertainty. 52

Figure 4.1: Feedthrough compensation using cross coupled capacitance. 55
Figure 4.2: Feedthrough cancellation using an attenuation buffer. 56
Figure 4.3: Proposed ‘double-switching’ architecture.

Figure 4.4: Illustrates the advantage of the double switching architecture. (a) Result of the conventional architecture. (b) The double switching architecture provides an equal voltage drop ($\Delta V_1 = \Delta V_2$) at the input of both switches.

Figure 4.5: Switched buffer and feedback buffer used in the double-switching architecture.

Figure 4.6: Schematic of the linear output buffer.

Figure 4.7: The simulated small signal gain and $S_{22}$ parameter of the 50Ω output buffer.

Figure 4.8: The simulated $IM_3$ of the output buffer.

Figure 4.9: Output waveform and spectrum of the high-resolution track-and-hold. $F_{in} = 1GHz/1.1GHz$ and $F_S = 40GS/s$.

Figure 4.10: Layout of the high resolution track-and-hold and output buffer.

Figure 5.1: Time interleaved architecture, consisting of 8 interleaves.

Figure 5.2: Wideband HBT buffer using gm-boosting.

Figure 5.3: The small signal model of the wideband buffer.

Figure 5.4: $IM_3$ comparison for the modified and simple architecture.

Figure 5.5: The small signal AC response of the modified and simple emitter follower.

Figure 5.6: The CMOS sample-and-hold architecture.

Figure 5.7: Schematic representation of the MOS track-and-hold.

Figure 5.8: Charge injection caused by a switching event.

Figure 5.9: Representation of the redistribution of charge. (a) Channel charged with $Q_{ch}$ when the switch is on. (b) Charge redistributes to the dummy devices when the switch is turned off.

Figure 5.10: Model of the switch in its ‘off’ state.

Figure 5.11: The NMOS track-and-hold.

Figure 5.12: The PMOS track-and-hold.

Figure 5.13: The gm-boosted buffer.

Figure 5.14: AC small signal bandwidth of the PMOS buffer.

Figure 5.15: $IM_3$ distortion in the PMOS gm-boosted buffer from 1-30GHz.

Figure 5.16: 50Ω pseudo-differential output buffer.

Figure 5.17: $IM_3$ distortion in the NMOS output buffer from 1-30GHz.

Figure 5.18: The simulated AC gain and $S_{22}$ parameter of the buffer.

Figure 5.19: The 8 phase clock generator.

Figure 5.20: The divide by 2 circuit.

Figure 5.21: Differential CML latch.

Figure 5.22: Differential CML buffer.

Figure 5.23: A CML to CMOS converter.

Figure 5.24: The 8 phases of the clock are generated from a 20GHz input.

Figure 5.25: The 8 phases of the clock are generated from a 20GHz input.
Figure 5.26: MOS sample-and-hold time domain waveforms for 600mVpp differential input envelope. Sample-and-hold spectrum for 2-tones shown in the time domain plot. 91

Figure 5.27: Layout of the time interleaved MOS sampling system. 92

Figure 5.28: Layout of the high-resolution track-and-hold and time interleaved CMOS sampling system. 93

Figure 6.1: Typical differential SEF driven by a transconductor. 95

Figure 6.2: Balun driving point impedance. (a) Differential driving point impedance. (b) Common mode driving point impedance. 96

Figure 6.3: A single-ended to differential SEF architecture using a balun as a passive ‘transconductor’. 98

Figure 6.4: Balun model including coupling capacitances. 98

Figure 6.5: A pair of emitter followers driven with a differential small signal. 99

Figure 6.6: A comparison of the simulated and calculated performance of the pseudo-differential circuit, where amplitude mismatch is introduced. 101

Figure 6.7: A balun simulated in HFSS. 101

Figure 6.8: S-parameters gathered from simulation of the balun. 102

Figure 6.9: A single-ended to differential SEF structure. 103

Figure 6.10: A differential to differential SEF architecture using a transformer as a passive transconductor. 103

Figure 6.11: A 1:1 transformer modeled in HFSS. 104

Figure 6.12: S-parameters gathered from simulation of the transformer. 105

Figure 6.13: A differential to differential transformer based SEF track-and-hold architecture. 105

Figure 6.14: Circuit schematic of the 50Ω output buffer. 106

Figure 6.15: Circuit schematic of the differential clock driver. 107

Figure 6.16: Simulated time domain waveform and output spectrum for the single-ended to differential SEF where $F_{in} = 5.1/5.2GHz$ and $F_s = 30GS/s$. 108

Figure 6.17: IIP3 plots simulated for the system, where $F_{in} = 5.1/5.2GHz$ and $F_s = 40GS/s$. 109

Figure 6.18: Layout of the balun driven track-and-hold chip designed in IBM 9HP. 111

Figure 6.19: Layout of the transformer driven track-and-hold chip designed in IBM 9HP. 111
LIST OF TABLES

Table 4.1: Compensated Track-and-Hold Performance Comparison . . . . . 65
Table 6.1: Track-and-Hold Performance Comparison . . . . . . . . 110
ACKNOWLEDGEMENTS

First and foremost I’d like to thank my advisor, Professor James Buckwalter. Without his support, my time as a Masters student would have been far less productive and enjoyable. Through his support I’ve been given access to processes and technologies that would otherwise have been impossible. I’ve thoroughly enjoyed our conversations and learned so much over the past year and a half. He has enabled me in so many ways and empowered me as a circuit designer. For that I am grateful.

I would also like to thank Professor Daniel Sievenpiper, who supported me during my first quarter as a graduate student. He always made himself very available to discuss new ideas, or resolve issues when they would arise. I always appreciated his intuitive and example-based approach to explaining electromagnetic concepts.

I want to thank my good friend and mentor Jeff Warner, for giving me the opportunity to prove myself as a young engineer. Without his dedication to teaching, positive motivation and humor I’d likely find myself traveling down a different road. I will always be very appreciative of his mentorship.

To my committee members, Professor Peter Asbeck and Dean Lawrence Larson, thank you for giving me your time. It’s been a privilege learning from some of the best. Additional thanks goes to Professor Asbeck for his time discussing HBT physics with me. Special thanks to Dr. Larson for personally giving me a tour of the mm-wave lab, which initially lead me to pursue circuit design.

To James Li at HRL Laboratories, thank you for providing me with simulation guidance and the image of an InP device. Our discussions were helpful.

In addition, I’d like to thank my lab mates for being such a fun and productive group. They’ve all been enjoyable to work with. I’d like to give special thanks to my lab mate Tim Gathman who collaborated with me on the projects in Chapters 4 & 5, and suggested the concept in Chapter 6. The work in Chapters 4 & 5 is included in
his Doctoral Dissertation entitled, “Highly-linear Sampling Receivers in Silicon and BiCMOS Processes for Multi-GS/s Optically and Electrically Sampled Systems.” I appreciate the time he invested to explain these concepts to me. Without him, the project wouldn’t have happened.
VITA

Santa Rosa Junior College 2005-2009

University of California, San Diego 2009-2011
Bachelor in Science, Electrical Engineering Cum Laude

University of California, San Diego 2011-2013
Master in Science, Electrical Engineering

PUBLICATIONS


FIELDS OF STUDY

Studies in Photoelectrochemical Processes, Device Fabrication and Electron Microscopy
Professor Deli Wang

Studies in Electromagnetics and Antennas
Professor Daniel F. Sievenpiper

Studies in High Speed Integrated Circuits
Professor James F. Buckwalter
BiCMOS Sampling Circuits for High-Speed Data Converters

by

Kristian Norgaard Madsen

Master of Science in Electrical Engineering (Electronic Circuits & Systems)

University of California, San Diego 2013

Professor James F. Buckwalter, Chair

This thesis introduces several high-speed, high-linearity sampling circuits designed in InP and SiGe BiCMOS processes, with applications in tests and measurement, optical networking and direct RF sampling. As a result of expanding high frequency operability, test and measurement equipment requires increased bandwidth capacity, which places increasing strain on data converters in these systems. In addition, optical backplanes have enabled 100Gbps data rates in high-speed network systems, pushing electrical transceivers to support sampling rates in excess of 56GS/s with moderate to high resolution. These applications place incredible pressure on broadband high-resolution data converter to perform at speeds not previously attained. As the demand for higher data rates continues to increase, the applicability and market for these systems will undoubtedly grow. A 40GS/s high-resolution track-and-hold, and time interleaved system are introduced for broadband Nyquist rate sampling in high bit rate optical sampling and measurement applications.
Direct RF sampling possesses advantages when used in reconfigurable radios, which address developing markets in wireless communications. This work introduces a pair of novel track-and-hold amplifiers capable of operating at 30-40GS/s, with low power consumption and high linearity. These systems are useful for direct RF sampling in reconfigurable radios, where power consumption is a limiting factor. The architectures described in this work help to advance the state of the art, through innovative circuit techniques which improve upon existing technologies, and support high bit rate and high frequency applications.
Introduction

This thesis discusses circuits and devices used in sampling systems for high-speed data conversion. These systems have applications in high frequency measurement, optical sampling and equalization, and direct digitization of RF signals. In particular, optical backplanes have enabled 100Gbps data rates, requiring electrical transceivers to support sampling rates in excess of 56GS/s with moderate to high resolution. Direct RF sampling possesses advantages when used in reconfigurable radios, opening up new avenues for wireless communications. As the demand for higher data rates continues to increase, the applicability and market for these systems will undoubtedly grow.

Chapters 1 & 2 cover introductory material regarding the devices and circuits which have enabled data conversion at 10’s of GS/s. Chapter 3 focuses specifically on the non-ideal circuit behavior which limits the resolution of these systems, and the following chapters describe the design of several sampling circuits useful in high-resolution, high-speed applications.
Chapter 1

Heterojunction Bipolar Transistors

The following chapter discusses the operation and design of bipolar transistors and leads into the topic of heterojunction devices. Heterojunction bipolar transistors have enabled drastic improvements in circuit performance where device speed is critical, playing an essential role in the advancement of start-of-the-art microwave and mixed-signal circuits. This chapter is an exploration of the science and design of modern HBT and BiCMOS processes, with an emphasis on mature high-speed technologies.

1.1 The Bipolar Junction Transistor

Bipolar devices rely on the injection and transport of minority carriers into and across the narrow base region in order to achieve current gain. The speed and efficiency with which the minority carriers are injected and transported across the base dictates the current gain and cut off frequency of the device. To model the carrier action in the BJT simply, we can consider 5 current components in an NPN configured device: \( I_{En} \), \( I_{Ep} \), \( I_R \), \( I_B \), and \( I_C \). This model assumes the base current is
completely comprised of holes, the collector current is entirely electrons and that no recombination occurs at either of the junctions.

Figure 1.1: Current components in an NPN Bipolar Junction Transistor (BJT).

The collector current $I_C$ is the injected electron current $I_{En}$ less the recombination current $I_R$ (Eq. 1.1). The base current is the sum of the injected hole current $I_{Ep}$ and the recombination current $I_R$ (Eq. 1.2).

$$I_C = I_{En} - I_R$$  \hspace{1cm} (1.1)
$$I_B = I_{Ep} + I_R$$  \hspace{1cm} (1.2)

The current gain of the device is defined as the ratio between the collector and base currents (Eq. 1.3).

$$\beta = \frac{I_C}{I_B} = \frac{I_{En} - I_R}{I_{Ep} + I_R}$$  \hspace{1cm} (1.3)

There are two fundamental efficiency parameters which dictate whether the device operates as a transistor and is capable of achieving current gain. First, the forward-biased junction between the base and emitter must be capable of supplying more electrons to the base than holes to the emitter. The efficiency of this transfer is quantified by the ratio of electron current to hole current across the junction (Eq. 1.4).

$$\gamma = \frac{I_{En}}{I_{Ep}}$$  \hspace{1cm} (1.4)
Second, once the electrons are injected they must be capable of traversing the base region without recombining before reaching the collector. The efficiency of this transport is quantified by the base transport factor shown in Eq. 1.5, which is simply the ratio of extracted electrons to injected electrons [1].

\[ \alpha = \frac{I_C}{I_{En}} \] (1.5)

Transistor action in the device can be modeled very simply using these two efficiency parameters. The reader should be warned that many more factors dictate the performance of bipolar devices; however, the described model provides a simple basis for understanding bipolar action. In addition, this simplified model allows us to contrast the operation of heterojunction and homojunction devices.

A simple manipulation of Eq. 1.3 yields Eq. 1.6, allowing a definition of current gain in terms of the device efficiencies, \( \gamma \) and \( \alpha \).

\[ \beta = \frac{\alpha \gamma}{1 + \gamma(1 - \alpha)} \] (1.6)

As recombination current diminishes, the base transport factor approaches unity and the current gain of the device tends toward the ratio \( \gamma \). This ratio establishes the maximum current gain \( \beta_{\text{max,BJT}} \), for the device. This can also be shown easily by removing the recombination current from Eq. 1.3.

\[ \beta_{\text{max,BJT}} = \frac{1\gamma}{1 + \gamma(0)} = \frac{I_{En}}{I_{Ep}} \] (1.7)

Numerical values can be derived from these expressions by considering the PN junction formed by the base and emitter. As in the case of a PN junction diode the
current across the base emitter junction is a function of the applied bias, as shown in Eq. 1.8 and Eq. 1.9.

\[ I_{En} = \frac{N_{E}D_{n,B}}{L_{n,B}}(e^{\frac{qV_{BE}}{kT}} - 1) \]  

\[ I_{Ep} = \frac{P_{B}D_{p,E}}{L_{p,E}}(e^{\frac{qV_{BE}}{kT}} - 1) \]

Given these equations the maximum current gain \( \beta_{\text{max},BJT} \) can be rewritten in terms of the intrinsic device parameters as in equation Eq 1.10. \( N_{E} \) and \( P_{B} \) are the number of electrons in the emitter and holes in the base, respectively. \( D_{n,B} / L_{n,B} \) and \( D_{p,E} / L_{p,E} \) are the diffusion rates / lifetimes of electrons in the base and holes in the emitter, respectively.

\[ \beta_{\text{max},BJT} = \frac{N_{E}D_{n,B}L_{p,E}}{P_{B}D_{p,E}L_{n,B}} \]  

The ratio \( D_{n,b}L_{p,e}/D_{p,e}L_{n,b} \) is generally modest and by itself does not constitute high current gain. In order for the device to achieve a large or even modest current gain, the emitter of the device must be more heavily doped than the base. As the doping in the emitter increases, more electrons are injected into the base than holes are injected into the emitter, improving the devices current gain. The naive device designer might quickly conclude that in order to produce the best transistor one ought to dope the emitter as heavily as possible and dope the base as lightly as possible. There are two immediately apparent reasons why this is not necessarily the best approach to take.

1. As the emitter is doped more heavily the width of the base-emitter space-charge region is reduced which subsequently increased the depletion capacitance of the junction. This increased has a direct impact on the speed of the device.

2. More importantly, the doping in the base region dictates its intrinsic resistance. As the doping in this region decreases, the resistance is increased, which
reduces device speed and increases base thermal noise.

Given these considerations the designer is faced with a trade off between current gain and device speed. There may be a sweet spot in the design of such a device which suits the needs of the application, but it’s clear that in the design of a homojunction device the designer is given a limited number of intrinsic device properties with which to tune its performance.

1.2 The Heterojunction

![Diagram of heterojunctions](image)

Figure 1.2: Three types of heterojunctions formed in semiconductors [2].

Heterojunctions are formed by either an abrupt or tapered junction of two dissimilar materials. The most important factor in semiconductor devices is the difference in conduction and valence band energies between the two materials. Fig. 1.2 illustrates the three types of heterojunctions which can be formed. A Type I junction Fig. 1.2a is formed by two materials in which the first material has a bandgap larger than the second, and the electron affinity is greater in the second than in the first. A Type II junction Fig. 1.2b is formed then the conduction and valence bands in the first material have higher potential than the conduction and valence bands in the second material respectively. In the case of the Type III junction Fig. 1.2c both the conduction and valence band in material 1 are at a higher potential than the valence band in material 2.
When the two semiconductors are brought into contact with each other a band
gap discontinuity is formed. Anderson’s rule is commonly used to describe the offset
in energy between the two materials. Eq. 1.11 & Eq. 1.12 describe the band offsets
between the two materials, where the offsets are quantified by deltas in the conduction
and valence band respectively [2].

\[ \Delta E_C = \chi_B - \chi_A \quad (1.11) \]
\[ \Delta E_V = (\chi_B + E_{GB}) - (\chi_A + E_{GA}) \quad (1.12) \]

A hypothetical system is depicted in Fig. 1.3 where material A is N-type and has a
wider bandgap than material B which is P-type. When the materials in this system
are brought into direct metallurgical contact (i.e. the material composition changes
abruptly at the junction) an abrupt bandgap discontinuity is generated and a Type
II heterojunction is formed, as shown in Fig. 1.4. The abrupt changes in the bands
are dictated by Anderson’s rule and the bending in the conduction and valence bands
can be determined by solving the Poisson equation [3]. This hypothetical system is a
reasonable model of the emitter-base junction for most modern Type II heterojunc-
tion bipolar transistors.
The current through an abrupt heterojunction with a spike in the conduction band is the product of thermionic and field emission. Thermionic emission occurs when the electron has sufficient thermal energy to overcome the potential barrier and field emission occurs when the electron has insufficient energy to overcome the barrier but is capable of tunneling through the barrier. These current components can be derived theoretically, but the derivation is rather cumbersome [2]. For the sake of ease it is assumed that the junction is compositionally graded in such a way that no band edge discontinuities are formed. This drastically reduces the difficulty of determining the current components, without diminishing the insight gained from analyzing heterojunction devices. The devices modeled in the following sections are represented with graded heterojunctions, which experience only thermionic emission currents across the junction.

1.3 Heterojunction Bipolar Transistors

In 1951 William Shockley was awarded a patent entitled, “Circuit Element Utilizing Semiconducting Material” in which he discussed the operation of a bipolar device and stipulated that added benefits would come by constructing the device
from dissimilar materials [4]. Shockley was the first to eluded to the usefulness of such a system, but by 1957 Herbert Kroemer elaborated on the idea by presenting a thorough analysis of the properties of carrier transfer across a heterojunction [5]. It wasn’t until the 1970’s that these junctions could be accurately produced [6] but Kroemer’s work provides the basis for analysing HBTs.

The HBT can be modeled with the same 5 current components used in the analysis of the BJT. Fig. 1.5 represents a device with a Type II junction formed between the emitter and base. In this case the device is modeled using band diagrams to illustrate the improvements made by using a wide bandgap emitter. The difference in valence band energies is much greater than the conduction band energies in this junction. Holes in the base see a larger potential barrier in the direction of the emitter than electrons in the emitter see in entering the base. This directly improves the ratio of injected electrons to injected holes.

The current components across the emitter base junction are a function of the intrinsic device properties and the energy barriers seen by the minority carriers.
These current flows are represented mathematically by Eq. 1.13 and Eq. 1.14.

\[ I_{En} = \frac{N_E D_{n,B}}{L_{n,B}} e^{-q V_n / kT} \]  \hspace{1cm} (1.13)

\[ I_{Ep} = \frac{P_B D_{p,E}}{L_{p,E}} e^{-q V_p / kT} \]  \hspace{1cm} (1.14)

The difference in barrier height is captured in Eq. 1.15 and is equivalent to the difference in bandgap between the two semiconductors.

\[ q(V_p - V_n) = \Delta E_G \]  \hspace{1cm} (1.15)

This leads to a new expression for the maximum current gain of the heterojunction device Eq. 1.16.

\[ \frac{I_{En}}{I_{Ep}} = \beta_{max, HBT} = \frac{N_E D_{n,B} L_{p,E}}{P_B D_{p,E} L_{n,B}} e^{\Delta E_G / kT} \]  \hspace{1cm} (1.16)

The maximum current gain of the device \( \beta_{max, HBT} \) is controlled by the same intrinsic properties as the BJT, in addition to the exponential term. As in the case of the BJT the intrinsic properties can be tuned marginally to improve device performance; however, for a heterojunction device the current gain becomes a strong function of \( \Delta E_G \) which produces an additional degree of freedom with which to tune the device performance. Because of this, HBT devices can achieve high current gain without excessive increases in base resistance or emitter-base depletion capacitance.

### 1.4 SiGe BiCMOS

Pure silicon is not an ideal semiconductor in the sense that it possesses relatively low carrier mobility and max carrier velocity. Because of these properties it is considered a “slow” semiconductor; however, silicon has the advantage of forming a native
oxide which is essential for producing MOS devices, which are necessary for low power logic applications. In addition silicon has the ability to support high levels of integration, which III-V materials are incapable of. The advantage of SiGe BiCMOS is the improvement to the Si BJT while maintaining the capacity to support high transistor counts, necessary in digital logic applications [7]. The following section discusses the design of SiGe HBTs.

Germanium (.66 eV) has a narrower bandgap than silicon (1.12 eV) and larger lattice constant. Introducing germanium to silicon produces compressive strain which further increases bandgap shrinkage in the alloyed material. For each 10% of germanium added to silicon there is a 75meV shrink in bandgap. The valence band is affected most heavily which makes the SiGe compound useful in npn HBTs. The mobility of SiGe is also higher than pure Si which improves its carrier transport ability. In addition SiGe films are limited to small thicknesses which necessitates its use in the
base region of the device. SiGe HBTs are generally constructed with an n-Si emitter p-SiGe base and n-Si collector producing a double heterojunction device. Although the first functional SiGe HBT had been demonstrated in 1987, it was an advancement in 1990 which bolstered a device with peak $f_t$ of 75GHz. The speed of this device was roughly double that of its mature Si counterparts [8].

SiGe-base HBTs benefit in a number of ways from their compound base. The base of the device is generally graded in such a manner that the germanium concentration increases linearly from the emitter-base junction to the collector of the device. This creates a narrowing of the bandgap in the base, effecting the conduction band primarily. This slope in the conduction band improves the transport of minority carriers across the base, reducing base transit time and improving device speed. This is coupled with the exponential increase in current gain which is derived from the germanium composition in the base which lowers the potential hill seen by electrons being injected from the emitter [7].

Fig. 1.6 illustrates the band diagram of the SiGe base HBT. Note that the solid line represents a standard Si BJT and the dashed line represents a device with the SiGe graded base. The $\Delta E_G$-Junction indicates that there is a narrowing in the conduction band at the base-emitter junction which improves the injection of electrons. $\Delta E_G$-Drift varies across the base starting with a minimum near the emitter-base junction. The sloped conduction band improves minority carrier drift in the base.

These advantages have placed SiGe HBTs in a position to contend with modern III-V devices, while being ahead of the best Si devices by at least a factor of 2 [9]. Chapter 4 discusses the design of ultra-high speed sampling systems in SiGe BiCMOS, using the IBM 9HP process. To demonstrate the high speed performance of these devices a simulated plot of $f_t$ and $f_{max}$ was included in Fig. 1.7. The HICUM model is used to represent the behavior of the SiGe HBTs in this library. The fastest of these
Figure 1.7: Simulated $f_t$ and $f_{max}$ curves from a range of devices sizes included in the IBM 9HP library.

devices show peak cutoff frequencies nearing 300GHz. 9HP incorporates these high speed HBTs alongside 90nm CMOS devices on the same substrate, enabling a large degree of design freedom and flexibility.

### 1.5 InP HBTs and an InP/Si BiCMOS Technology

InP posses a number of properties which gives it an advantage over GaAs. The first advantage to using InP is that it has a higher peak carrier velocity which improves its high frequency behavior. The base is generally engineered to produce a high drift field which sweeps electrons quickly across the region. Additionally, when the collector is biased in active mode even higher field strengths can exist in
this region. Where high drift fields exist, carriers can attain higher velocities which are bounded by the peak carrier velocity of the material. In InP, where this peak value is substantially greater than GaAs, higher velocities can be achieved leading to overall decreases in base and collector transit times and increased high frequency performance. The second advantage is that InP devices also have lower turn on voltages which makes them desirable for applications where power consumption is critical [9].

InP devices are generally constructed with a wide bandgap InP emitter and InGaAs base. They manifest themselves in two forms, as either SHBTs or DHBTs. Single heterojunction (SHBT) devices are formed with only one heterojunction, which exists between the emitter and base of the device. The collector maintains the same material composition as the base but is doped to the designers specifications. Double heterojunction (DHBT) devices are formed with two heterojunctions, one between the emitter and base and the other between the base and collector. These are generally more common than single heterojunction devices, because they bolster both higher breakdown voltages and output impedance. InP devices generally outshine GaAs devices in frequency performance, but fall short on their power handling capabilities [2]. The first InP/InGaAs HBT lattice matched to InP, with a carbon doped base was reported in 1992 and showed DC current gain above 500 [10].

Chapters 5 and 6 discuss ultra-high speed sampling circuits designed in an InP/Si BiCMOS technology developed and licensed by HRL Laboratories, LLC [11, 12]. The technology uses heterogeneous integration to fabricate InP HBTs alongside standard Si CMOS. The technology was developed as a platform for the integration of high density digital CMOS circuitry alongside ultra high-speed III-V devices. The process integrates 250nm emitter width HBTs on top of IBM’s 90nm bulk CMOS process. The InP HBT devices are transferred and bonded to the top level of the
oxide produced by the IBM process, using a proprietary technique [13]. Vias are formed to connect the high-speed devices to the underlying CMOS circuitry. The general concept is captured in Fig. 1.8 along with an image of the InP HBT shown in Fig. 1.9.

Figure 1.8: A cross sectional diagram of the InP/Si BiCMOS stack up, reproduced from [13].

Figure 1.9: Image of an InP DHBT from the G4 library. © 2010 HRL Laboratories, LLC. All Rights Reserved. Reproduced with permission from [14].

To demonstrate the high speed capabilities of the InP devices, simulated $f_t$
and $f_{\text{max}}$ plot are included in Fig. 1.10. The VBIC model is used to represent the behavior of the InP HBTs in this library [15]. The fastest device in the library possesses a maximum cut off frequency near 330GHz and maximum oscillation frequency of roughly 310GHz. Bonding these devices above the top metal layers has the added benefit of reducing device and routing parasitics, which are intrinsic to substrate-level devices. The drawback however, is the thermal isolation provided by the oxide surrounding the HBTs. To combat this the InP devices include a thermal via to wick heat away from the devices and into the substrate.

Figure 1.10: Simulated $f_t$ and $f_{\text{max}}$ curves for some of the InP devices in the BiCMOS library.
1.6 A Technology Comparison

The following section uses device speed to make a number of comparisons between InP and SiGe BiCMOS processes, both used in this work and in recent development. To demonstrate the performance of the fastest devices from both the InP and SiGe libraries used in this work, $f_t$ and $f_{max}$ have been plotted versus emitter current density. As Fig. 1.11 demonstrates, the SiGe HBTs require roughly 3 times higher emitter current density to achieve peak $f_t$ while requiring almost an order of magnitude higher current density to achieve peak $f_{max}$. It is clear from this comparison that the SiGe devices require much higher power density to achieve their peak operating speed, when compared with the InP process.

![Figure 1.11: Comparison of $f_t$ and $f_{max}$ versus emitter current density, for the InP and SiGe BiCMOS processes.](image)
The reader might question why this is the case. To first order, $f_t$ can be approximated by the ratio $g_m/c_{be}$, which suggests that the SiGe device has higher $c_{be}$ per area, and thus requires more current density to achieve peak $f_t$. Higher $c_{be}$ is likely due to a smaller $\Delta E_G$ in the SiGe devices, which increases the amount of emitter doping that’s required in these HBTs. Higher emitter doping increases depletion capacitance in the base-emitter junction, and makes for a noticeable difference in emitter current density between the SiGe and InP devices.

![Figure 1.12: Device speed comparisons for SiGe HBTs over the past 10 years.](image)

In addition, a quick overview is captured in graphical form which highlights recent technological advancements in SiGe and InP HBTs over the past 10 years. Two plots are included as summaries of device speeds reported in mainstream technical
literature. Fig. 1.12 captures the rise in SiGe HBT device speed since 2001. Included in this plot, are results reported for HBT devices developed as isolated bipolar technologies and HBT devices incorporated in BiCMOS technologies. There is no clear distinction between the two. Fig. 1.13 shows the development of InP HBTs, including both SHBT and DHBT designs. It is clear from the plot that DHBT designs are favored, as they greatly outnumber reported results for SHBTs, and tend to provide higher cutoff and max oscillation frequencies. The fastest DHBT reported a maximum oscillation frequencies of 1.1THz, while maintaining a cutoff frequency of 520GHz [16]. This 130nm InP DHBT has a 25nm thick carbon doped InGaAs base, and requires an emitter current density of 10mA/µm² to achieve peak high frequency performance.

Figure 1.13: Device speed comparisons for InP HBTs over the past 10 years.
Chapter 2

Sampling Circuits and Systems

The following chapter is meant to familiarize the reader with a variety of circuit blocks and sampling systems which provide a basis for the material covered in later chapters. First the concept of sampling is introduced, leading into a discussion on track-and-hold and sample-and-hold operation. Finally the concept of time interleaving is introduced, along with a brief glimpse at the benefits and drawbacks of these systems.

2.1 Sampling Theory

At its fundament, the word, ‘sampling’ refers to the use of a finite subset of a larger or potentially infinite set. In the electrical domain this refers to capturing information at discrete points from a continuum of supplied information. This process is represented below for a continuum of supplied voltages. Fig. 2.1a represents the continuous signal presented to the sampling system. These voltages are analog and contain amplitude information at all points in time. The continuous signal is then overlapped with discrete timing instants which illustrate the times at which the signal
is sampled and the amplitude of the signal is recorded (Fig. 2.1b). The result is a
discrete time signal shown in Fig. 2.1c which contains amplitude information only at
discrete points in time. The discrete signal contains information about the original
signal only at points where the signal was sampled. This results in a reduction in the
information retained relative to the information provided.

2.2 Sampling Circuits

There are two fundamental circuit blocks which are used in data conversion
systems and therefore deserve some attention. The first is commonly referred to as a
track-and-hold, which can be used to implement the second circuit called a sample-
and-hold.
2.2.1 Track-and-Hold

A track-and-hold is an elementary sampling circuit which produces a zero-order hold response. Unlike the discrete time signal shown in Fig. 2.1c, the zero-order hold contains amplitude information in continuous time. This is an unavoidable outcome of implementing sampling systems with physical circuits which are inherently bandwidth limited. The track-and-hold passes the input signal to the output during the track mode and retains the last input value observed during the hold mode. This action is represented in Fig. 2.2.

![Diagram of track-and-hold operation]

Figure 2.2: Operation of the track-and-hold, representing the continuous input signal, clock phases and output signal.
2.2.2 Sample-and-Hold

A sample-and-hold can be generated in a number of ways, two of which are shown in Fig. 2.3. The first is a cascaded architecture in which the two track-and-holds are clocked complementary to one another. The second incorporates two parallel track-and-holds which are again clocked with complementary phases, but are instead fed into an analog multiplexer. The multiplexer is switched such that the hold phase of each THA appears at its output. The resulting output of both systems is a waveform looking like a series of steps, which alter their values only at the edges of the clock.

![Diagram](image)

Figure 2.3: Two methods of producing a sample-and-hold. (a) Formed by cascading two track-and-holds. (b) Formed by two parallel track-and-holds fed into an analog multiplexer.

The time domain waveform at the output of an ideal sample-and-hold is shown in Fig. 2.4. This is an idealization of a sample-and-hold system which requires no acquisition time in either the sample-to-hold or hold-to-sample transition in the secondary track-and-hold. In actuality there would be a rise or fall time between the hold periods.
Figure 2.4: Operation of the sample-and-hold. The output of the single track-and-hold is shown in dashed lines.

Figure 2.5: Operation of the sample-and-hold modeled with an impulse train and rectangular impulse response.

The sample-and-hold response can be modeled simply by a system which mixes the continuous time signal with an impulse train and convolves the result with a unity rectangular pulse having a duration as long as the sampling period. This concept is illustrated in Fig. 2.5 and allows us to derive a mathematical transfer function for the
sample-and-hold. The resulting transfer function is shown in Eq. 2.1 where the input is phase shifted by \( e^{-j\pi fT} \) and attenuated in amplitude by the sinc response [17].

\[
H(j\omega) = e^{-j\pi fT} \left[ \frac{\sin(\pi fT)}{\pi fT} \right]
\]  
(2.1)

### 2.3 Time Interleaved Sampling

Time interleaved sampling is a process used to sample information in parallel on several offset phases to increase the effective sampling rate of the system. Time interleaving helps to break the trade off between bandwidth and resolution in applications where the required bandwidth and resolution exceed the capabilities of a single analog-to-digital converter (ADC). Time interleaving systems employ multiple parallel ADCs operating at lower frequency to reconstruct a higher bandwidth signal. As Fig. 2.6 suggests, the input is sampled by ‘n’ parallel ADCs with ‘n’ phases encompassing the unit circle.

![Figure 2.6: Block diagram of a time interleaved sampling system with associated phase relationships.](image)
2.3.1 4-Phase Time Interleaved Track-and-Hold

In what remains of this section, a simulated 4-phase track-and-hold is used to illustrate the concept of time interleaving. This system differs slightly from the concept illustrated in Fig. 2.6, in that it consists only of simple sampling circuits and that no quantization occurs. The simulation remains valuable in illustrating the general concept of time interleaved systems. In reality, each parallel branch would likely consist of a complete ADC. The general concept of time interleaving and issues which plague these systems can easily be conceptualized in the context of a larger or more complex data conversion system, once this simple illustration is understood. The system used to exemplify these issues is shown in Fig. 2.7.

![Diagram](image)

Figure 2.7: A 4-phase time interleaved track-and-hold proposed as an example for the following section.

The system consists of 4 parallel track-and-holds which are clocked with phases $0^\circ$, $90^\circ$, $180^\circ$ and $270^\circ$, respectively. Each clock phase has a 50% duty cycle and has a frequency $F_s$. Fig. 2.8 shows a simulated signal sampled by the 4-phase time
interleaved track-and-hold, with phases corresponding to the diagram. The time
domain waveform shows the overlapped output of the 4 track-and-holds. Below the

![Relative Amplitude (V) vs. Relative Time (T_s)](image1)

![Relative Amplitude (dB) vs. Relative Frequency (F_s)](image2)

Figure 2.8: Time domain plot of the 4 overlapping track-and-hold phases, along with
the output spectrum of a single track-and-hold.

waveforms is the associated output spectrum of a single track-and-hold. It’s clear
from this plot that the input spectrum is aliased around $F_s$. This appears to limit
the bandwidth of the system to $F_s/2$, because spectral content above the Nyquist
frequency of the track-and-hold is folded down. Although this is true for a single
track-and-hold clocked at $F_s$, the time interleaved system bandwidth can extend to
$n \cdot F_s/2$ without the occurrence of harmful aliasing.
2.3.2 Signal Reconstruction

If the sampled signals are ‘stitched’ back together by means of a multiplexer as shown in Fig. 2.9, the resulting waveform has a zero-order hold response at four times the sampling rate of the individual interleaved cells, which in this case is $4F_s$. The simulated response of the system is shown in Fig. 2.10 along with the associated output spectra. In this simulation a tone is applied with a frequency in excess of $F_s/2$. The first spectrum represents the output of a single track-and-hold, where aliasing produces unwanted tones in the band of interest. The second spectrum represents the reconstructed signal. This demonstrates that the reconstructed signal contains aliased versions of the signal only around the effective Nyquist rate $4F_s$. The process of ‘stitching’ the signal together at a rate ‘n’ times the sampling frequency, produces an effective Nyquist bandwidth of $2F_s$ from parallel interleaved units with lower Nyquist bandwidths of $F_s/2$. 

Figure 2.9: A 4-phase time interleaved track-and-hold with a multiplexed output. The effective sampling rate increases to $4F_s$. 
Figure 2.10: Time domain plot of the reconstructed signal, along with the spectrum of a single track-and-hold and reconstructed signal.

From this demonstration it is clear that the bandwidth of the system can be increased by a factor ‘n’ when time interleaved sampling units or ADCs are used in parallel. If an ADC is designed to operate at a specific rate with a given resolution, ‘n’ interleaved versions of the ADC should ideally operate at n-times the sampling rate, with equivalent resolution. There are several issues which arise from interleaving, and are discussed in the following subsections.
2.3.3 Gain Mismatch

When time interleaved systems are implemented on physical substrates, process deviation, non uniformities and physical layout differences lead to device and circuit mismatch. Consequently gain mismatch, offset error and timing skew occur between parallel circuits. The following subsections present simple models to capture the affects of these circuit non-idealities. To avoid lengthy derivation, simulations are accompanied by intuitive explanations to provide reasoning for the resulting output errors.

Gain mismatch can be modeled with a gain element in the signal path of each parallel interleaved branch, which has a value $G$ plus a deviation $\Delta_x$ from its neighbors. This is shown in Fig. 2.11. The result of this gain error is the addition of spurious tones around multiples of the interleaved sampling rate $F_s$. For an input with frequency $F_{in}$, spurious tones appear at $n \cdot F_s \pm F_{in}$ where $n = 1, 2, 3 \ldots \infty$. The voltage error incurred by the gain mismatch has a periodic amplitude dependence on the input, which appears to be sampled at the rate $F_s$. This results in spurious
tones at the input frequency, which are then aliased around multiples of $F_s$. This is demonstrated by the output spectrum shown in Fig. 2.12. This spurious distortion is dependent on the amplitude of the input, and therefore limits the SFDR of the time interleaved system.

2.3.4 Voltage Offset

Time interleaved systems also suffer from constant voltage offsets between parallel channels. A system with voltage offsets is modeled in Fig. 2.13 where $\delta_x$ represents the additive offset incurred on each signal path. The result is the addition of spurious tones at the sampling rate $F_s$ and its multiples $n \cdot F_s$ where $n = 2,3,4 \ldots \infty$. A simulation demonstrates the generation of spurious tones for a four channel interleaved system, where voltage offsets between channels have a normal distribution with standard deviation of 10mV. The signal amplitude in this simulation was 1Vpeak.

The voltage offsets in this simulation are rather large in order to clearly demonstrate the effect of the nonideality. A physical system would likely incur much less
distortion as an outcome of offset. None the less, the results of the simulation are shown in Fig. 2.14 where the offsets spurs are highlighted with circles. As mentioned previously, these spurs occur at multiples of the sampling frequency. The spurs are
based on a fixed voltage offset; thus, unlike a nonlinearity, they do not grow/shrink with applied signal power. Because of this, the distortion incurred by channel offsets restricts the dynamic range of the system.

### 2.3.5 Systematic Timing Errors

Systematic timing errors affects the resolution of time interleaved systems, like gain mismatch and voltage offset. Systematic timing errors are static deviations from the ideal clock phases presented to each of the parallel sampling blocks. This error differs from aperture error and sampling jitter (discussed in Chapter 3), as it does not result from nonlinearities or noise, but is likely the result of unequal delays between the clock drivers and sampling switches, or mismatches in the clock drivers themselves. This error is illustrated in Fig. 2.15 for a system where acquisition time is shifted by a timing offset $\sigma_x$. This simple model ignores timing errors incurred when reconstructing the signal at the back end of the system.

![Figure 2.15: Model of a time interleaved system where systematic timing error is present between channels.](image-url)
For an input frequency $F_{in}$, systematic timing error results in output spurs at $n \cdot F_s \pm F_{in}$ where $n = 1, 2, 3 \ldots \infty$. Like gain mismatch, timing error results in periodic error voltages with dependence on the input frequency and appear to be sampled at the rate $F_s$. Thus, spurs are generated at the input frequency and aliased around multiples of $F_s$. Unlike gain mismatch, the error dependence is not on the amplitude of the input signal, but on its derivative. As the input frequency increases, so does the resulting distortion. This places a frequency dependent SFDR limitation on the interleaved system. A simulation of the 4-phase track-and-hold system was run, including these timing errors. The errors were chosen to match a standard random distribution, unlike the normal distributions used to simulate gain mismatch and offset error. The results of the simulation are shown in Fig. 2.16 where 3 tones were applied. The three tones are labeled ‘low’, ‘medium’ and ‘high’ as they correspond to increases in frequency. As can be seen from the resulting output spectrum, the error spurs grow with the input frequency which matches the conclusion stated above.

![Output error incurred by clock timing errors, shown for 3 input tones.](image)

Figure 2.16: Output error incurred by clock timing errors, shown for 3 input tones.
These errors can be partially compensated for using good layout and device matching techniques. These techniques ensure that clock lines are routed to provide equal delay and that devices match their neighbors accurately. These good practices don’t remove these errors completely; therefore, calibration can be used to compensate for these errors using built in algorithms. In addition, intelligent circuit topologies can be implemented to reduce the need for highly complex calibration systems. Chapter 5 discusses a circuit topology which can be implemented to reduce the affect of timing skew error in addition to decreasing the affect of jitter on sample acquisition.
Chapter 3

Track-and-Hold Design

The HBT track-and-hold is the fundamental building block around which most of following systems are built. For this reason the track-and-hold deserves a bit of explanation in order to reveal certain intricacies involved in its design. To develop the necessary background Chapter 3 focuses on the operation and nonidealities of switched emitter follower track-and-holds.

3.1 Switched Emitter Follower Architecture

The primary means of creating a sampling switch in a bipolar technology is through the use of a switched emitter follower (SEF). The SEF consists of a switching differential pair which turns an emitter follower on and off. In its ‘on’ state the emitter follower acts as a voltage buffer, transferring the input voltage signal onto a hold capacitor. In its ‘off’ state the emitter follower acts as an open circuit and the previously held voltage is maintained. It is with this action that a track-and-hold circuit is formed.
Figure 3.1: The switched emitter follower architecture.

The track mode of the SEF is shown in Fig. 3.2a when the device $Q_1$ forms the emitter follower, buffering the input to the output. In hold mode (Fig. 3.2b) device $Q_3$ is turned on, while $Q_2$ is turned off. The current is switched across the resistor $R$ which pulls the base voltage of $Q_1$ down. The voltage drop across $R$ must be great enough to turn off $Q_1$. It is with this in mind that $R$ is sized.

Figure 3.2: Operation of a switched emitter follower track-and-hold. (a) SEF in track-mode. (b) SEF in hold-mode.
3.2 Emitter Follower Linearity Analysis

The linearity of the track-and-hold is a very important design consideration when the circuit is to be employed in a high resolution data converter. Volterra series is employed to analyze the linearity of the emitter follower during track mode. As Fig. 3.3 demonstrates, the emitter follower is loaded by the hold capacitor at its output which contributes most significantly to the loading on the follower as the input frequency is increased. The small signal model in Fig. 3.4 represents the device input impedance, a linear output capacitance and a third order non-linearity in transconductance.

![Figure 3.3: Emitter follower driving a capacitive load, representing the SEF during track mode.](image)

![Figure 3.4: Non-linear small signal model of the capacitively loaded emitter follower.](image)
Volterra series describes a nonlinear system which is excited by a small amplitude signal. A comparison can be made to the use of Taylor series in describing an analytic function. As the input amplitude grows, so does the number of terms in the Volterra series. At a point the series diverges as in the case of the Taylor series. This approach is useful for systems with only a few nonlinearities which makes this method ideal for our purposes. The purpose of this exercise is to utilize block diagrams to model the nonlinearity and frequency dependent behavior in a way that improves intuition about these phenomenon [18].

The Volterra series representation of a nonlinear system is the sum of the outputs of the first n-order Volterra operators, as shown in Eq. 3.1. Each Volterra operator is the Laplace transform of the respective Volterra kernel. The Volterra kernels will not be used in this analysis, in favor of using their transforms. Each Volterra operator can be derived from the included block diagrams, which consist of the linear subsystems and their respective transforms.

\[ v_{out}(t) = H_1[v_{in}(t)] + H_2[v_{in}(t)] + H_3[v_{in}(t)] \]  

(3.1)

The block diagram shown in Fig. 3.5 represents the small signal model in Fig. 3.4 as a feedback system where \( z_{in} \) is considered infinite. The blocks \( G_1, G_2 \) and \( G_3 \) transform the voltage \( v_{be} \) into the current \( i_{out} \), which flows across the impedance \( Z_1 \). This current generates the output voltage \( v_{out} \) which is compared with the input voltage to generate the voltage \( v_{be} \), thus closing the loop. The blocks \( G_1, G_2 \) and \( G_3 \) represent the three transconductance terms, and \( Z_1 \) represents the capacitive impedance seen at the output of the emitter follower.
First we consider the factors which contribute linearly to the system. The first order operator is generated by an input signal which passes into the first order transconductance term $G_1$ to generate a voltage on the capacitor, dictated by $Z_1$. This voltage is compared with the input signal to complete the loop and reduce the output signal by $1 + T$, where $T$ is the loop gain of the system. In this analysis $R(\omega_x)$ is used to represent the gain reduction imposed by the loop, as in Eq. 3.3. The product of these terms results in Eq. 3.2 which represents the complete first order Volterra operator.
\[ H_1(\omega_1) = G_1 \times Z_1(\omega_1) \times R(\omega_1) \]  
\[ R(\omega) = \frac{1}{1 + G_1 \times Z_1(\omega)} \]  
\[ H_1(\omega_1) = \frac{G_1 Z_1(\omega_1)}{1 + G_1 Z_1(\omega_1)} \]

Next comes a derivation of the second order operator, which is a product of the second order nonlinear transconductance shown in Fig. 3.7. The error voltage \( v_{be} \) excites the second order transconductance nonlinearity to generate an output current which is the product of two frequencies \( \omega_1 \) and \( \omega_2 \). The second order nonlinearity is evaluated as the product of the two error signals \( R(\omega_1) \) and \( R(\omega_2) \) and the second order transconductance \( G_2 \). The current generated by the second order transconductance forms a voltage on the capacitor, which is a function of the two frequencies. Again feedback is applied, represented by \( R(\omega_1 + \omega_2) \) and the complete operator can be determined.

Figure 3.7: Second-order loop model.
It should be noted that the ordering of the terms $\omega_1$ and $\omega_2$ sometimes has an effect on the output of the operator. Eq. 3.5 represents the complete Volterra operator $H_2$ as the weighted sum of two terms, which represent all iterations of the two frequencies.

$$H_2(\omega_1, \omega_2) = \frac{1}{2} H'_2(\omega_1, \omega_2) + \frac{1}{2} H'_2(\omega_2, \omega_1) \quad (3.5)$$

A general solution for each of the iteration is represented by $H'_2(\omega_a, \omega_b)$ in Eq. 3.6. It is in fact true that both iterations are equal as shown in Equation 3.7.

$$H'_2(\omega_a, \omega_b) = G_2 \times R(\omega_a) \times R(\omega_b) \times Z(\omega_a + \omega_b) \times R(\omega_a + \omega_b) \quad (3.6)$$

$$H'_2(\omega_a, \omega_b) = H'_2(\omega_b, \omega_a) \quad (3.7)$$

This allows for a simplification of the second-order operator to the form shown in Eq. 3.8 & Eq. 3.9.

$$H_2(\omega_1, \omega_2) = H'_2(\omega_1, \omega_2) = G_2 \times R(\omega_1) \times R(\omega_2) \times Z(\omega_1 + \omega_2) \times R(\omega_1 + \omega_2) \quad (3.8)$$
To complete the analysis, the third-order operator is derived from two mechanisms which generate third-order nonlinearities. The third order kernel of the Volterra Series is generated in two ways. First by means of the third order non-linearity of the transconductor (Fig. 3.9) and through the second order non-linearity of the transconductor interacting with itself (Fig. 3.10). The third-order nonlinearity generated by the third and second order transconductance terms are denoted \( H_{3,1} \) and \( H_{3,2} \) respectively.

The contribution of the third order nonlinearity is represented by block diagrams in Fig. 3.9a and 3.9b.

\[
H_2(\omega_1, \omega_2) = \frac{G_2 Z_1(\omega_1 + \omega_2)}{(1 + G_1 Z_1(\omega_1))(1 + G_1 Z_1(\omega_2))(1 + G_1 Z_1(\omega_1 + \omega_2))} \tag{3.9}
\]

The error voltage excites the third-order transconductance nonlinearity producing three frequency components \( \omega_1, \omega_2 \) and \( \omega_3 \), which are then multiplied by \( G_3 \). This current interacts with the capacitance at the output to generate a voltage which
is a function of these three frequencies. The loop is closed and $R(\omega_1 + \omega_2 + \omega_3)$ is included as the feedback interacts with the gain of the system. Generation of the third order kernel by means of the third order non-linearity is described by Eq. 3.10.

$$H_{3,1}(\omega_1, \omega_2, \omega_3) = G_3 \times R(\omega_1) \times R(\omega_2) \times R(\omega_3) \times Z_1(\omega_1 + \omega_2 + \omega_3) \times R(\omega_1 + \omega_2 + \omega_3)$$

(3.10)

The second order nonlinearity of the transconductor has the ability to generate higher order terms when is interacts with itself. This process is shown in Fig. 3.10a & Fig. 3.10b. The result of this interaction is shown in Eq. 3.11.

![Figure 3.10: Models representing $H_{3,2}$.](image)

(a) Third-order loop model incorporating $G_2$.

(b) Third-order term generated by the second order nonlinearity.
\[ H_{3,2}(\omega_1, \omega_2, \omega_3) = -2G_2^2 \times R(\omega_1) \times R(\omega_2) \times Z_1(\omega_1 + \omega_2) \times R(\omega_1 + \omega_2) \times R(\omega_3) \times Z_1(\omega_1 + \omega_2 + \omega_3) \times R(\omega_1 + \omega_2 + \omega_3) \] (3.11)

As in the case of the second-order operator the ordering of the terms \( \omega_1 \), \( \omega_2 \) and \( \omega_3 \) bares significance on the result of the third-order operator. The third-order operator is the sum of six terms which represent the six iterations of the three frequencies, as shown in Eq. 3.12.

\[ H_3(\omega_1, \omega_2, \omega_3) = \frac{1}{6} H'_3(\omega_1, \omega_2, \omega_3) + \frac{1}{6} H'_3(\omega_1, \omega_3, \omega_2) + \frac{1}{6} H'_3(\omega_2, \omega_1, \omega_3) \]
\[ + \frac{1}{6} H'_3(\omega_2, \omega_3, \omega_1) + \frac{1}{6} H'_3(\omega_3, \omega_2, \omega_1) + \frac{1}{6} H'_3(\omega_3, \omega_1, \omega_2) \] (3.12)

Each iteration denoted by \( H'_3 \) consists of a term produced by the third order non-linearity, denoted \( H'_{3,1} \) and a term generated by the second order non-linearity, denoted \( H'_{3,2} \) shown in Eq. 3.13. Each individual iteration bares a superscript (’') in order not to be confused with the complete Volterra operator.

\[ H'_3(\omega_a, \omega_b, \omega_c) = H'_{3,1}(\omega_a, \omega_b, \omega_c) + H'_{3,2}(\omega_a, \omega_b, \omega_c) \] (3.13)

The sum of the six terms shown in Equation 3.13 can be simplified because some of the orderings produce equivalent solutions. All of the terms generated by the third order nonlinearity are equivalent, and the equal terms generated by the second order nonlinearity are shown in Eq. 3.14, 3.15 & 3.16.

\[ H'_3(\omega_a, \omega_b, \omega_c) = H'_3(\omega_b, \omega_a, \omega_c) \] (3.14)
\[ H'_3(\omega_a, \omega_c, \omega_b) = H'_3(\omega_c, \omega_a, \omega_b) \] (3.15)
\[ H'_3(\omega_b, \omega_c, \omega_a) = H'_3(\omega_c, \omega_b, \omega_a) \] (3.16)
In consideration of these simplifications the complete third order Volterra operator can be determined and is shown in Eq. 3.17.

\[
H_3(\omega_1, \omega_2, \omega_3) = \frac{Z_1(\omega_1 + \omega_2 + \omega_3)}{(1+G_1Z_1(\omega_1))(1+G_1Z_1(\omega_2))(1+G_1Z_1(\omega_3))} \times \left( G_3 - 2G_2^2 \left( \frac{1}{3} \frac{Z_1(\omega_1 + \omega_2)}{(1+G_1Z_1(\omega_1 + \omega_2))) + \frac{1}{3} \frac{Z_1(\omega_2 + \omega_3)}{(1+G_1Z_1(\omega_2 + \omega_3))} + \frac{1}{3} \frac{Z_1(\omega_1 + \omega_3)}{(1+G_1Z_1(\omega_1 + \omega_3))} \right) \right)
\]  

(3.17)

The purpose of deriving each Volterra operator was to provide a theoretical means for approximating the distortion of the emitter follower. For a system excited by a sinusoid \( v_{in} = A \cos(\omega_o t) \) the output is determined by summing the responses of the three Volterra operators shown in Eq. 3.18.

\[
y(t) = A \Re(H_1(j\omega_o)e^{j\omega_o t}) + \frac{A^2}{2} \Re(H_2(j\omega_o, j\omega_o)e^{j2\omega_o t}) + \frac{A^2}{2} \Re(H_2(j\omega_o, -j\omega_o)) + \frac{A^3}{4} \Re(H_3(j\omega_o, j\omega_o, j\omega_o)e^{j3\omega_o t}) + \frac{3A^3}{4} \Re(H_2(j\omega_o, j\omega_o, -j\omega_o)e^{j\omega_o t})
\]

(3.18)

To determine the effect of harmonic distortion the products generated at the harmonic frequencies are weighted against the signal at the fundamental. The ratio of the second and third order distortion products to the fundamental are referred to as \( \text{HD}_2 \) and \( \text{HD}_3 \) respectively. The result of these ratios are found in Eq. 3.19 & Eq. 3.20.

\[
\text{HD}_2 = \frac{\frac{A^2}{2} \Re(H_2(j\omega_o, j\omega_o))}{\Re(H_1(j\omega_o)) + \frac{3A^3}{4} \Re(H_3(j\omega_o, j\omega_o, -j\omega_o))} \approx \frac{A}{2} \left| \frac{H_2(j\omega_o, j\omega_o)}{H_1(j\omega_o)} \right|  
\]

(3.19)

\[
\text{HD}_3 = \frac{\frac{A^3}{4} \Re(H_3(j\omega_o, j\omega_o, j\omega_o))}{\Re(H_1(j\omega_o)) + \frac{3A^3}{4} \Re(H_3(j\omega_o, j\omega_o, -j\omega_o))} \approx \frac{A^2}{4} \left| \frac{H_3(j\omega_o, j\omega_o, j\omega_o)}{H_1(j\omega_o)} \right| 
\]

(3.20)

To simplify, the contribution of the third order nonlinearity to the fundamental tone
can be disregarded. This leads to the simplifications made in Eq. 3.19 & 3.20. For an HBT device the individual transconductance terms can be generated from a known collector current. Equations 3.21 & 3.22 are the result of the simplified expressions for HD$_2$ and HD$_3$ where circuit specific parameters, $C_H$, $I_{Bias}$, $f_{in}$, $V_{pk}$, and $V_t$ are used to quantify the theoretical linearity limitation of the SEF.

$$HD_2 = \left| \frac{V_{pk} V_t}{4} \left( \frac{2\pi f_{in} C_H}{I_{Bias}} \right)^2 \right|$$

(3.21)

$$HD_3 = \left| \frac{V_{pk}^2 V_t}{12} \left( \frac{2\pi f_{in} C_H}{I_{Bias}} \right)^3 \right|$$

(3.22)

Generally the second order distortion is disregarded in the differential SEF architecture, such that HD$_3$ limits the system’s performance. These expressions are derived analytically in other works [19].

A plot of these theoretical results is compared with a simulation of a simple emitter follower biased with 1mA, driving a 100fF capacitance at 1GHz. Fig. 3.11 shows that there is a fair amount of agreement between the simulated result, and the result derived from Volterra analysis. This analysis is useful for analyzing weak nonlinearities, like the transconductance nonlinearity of the emitter follower. The result provides information about the basic linearity limitations of this open-loop bipolar track-and-hold architecture, such that circuit parameters can be chosen to meet the minimum resolution requirements of the system.
Figure 3.11: A comparison between simulated and theoretical distortion in the SEF.

3.3 Droop Induced Non-Linearity

Droop is an issue in bipolar track-and-holds which is a product of the devices finite input impedance. In an ideal track-and-hold the voltage sampled on the hold capacitor is maintained at a fixed value during the hold phase; however, the held voltage is generally buffered to other circuits in the system, which introduces the cause of error. The finite input impedance of the buffer results in the removal of charge during the hold phase. As charge is removed from the hold capacitors their voltage is altered, resulting in voltage droop. This voltage alteration manifests itself as both common-mode and differential droop.
This voltage distortion is largely a common mode deviation, which can be somewhat negated in differential architectures, but not entirely. The input impedance of the bipolar buffer is a non-linear function of the applied voltage, which implies that charge is removed at unequal rates from either of the hold capacitors. This non-linear input impedance is attributed to $V_{be}$ induced $\beta$ variations in the two legs of the buffer [20].

### 3.4 Pedestal Error

Pedestal error refers to an error voltage sampled on the hold capacitor, which is due to the switching event [21]. This error arises from a number of factors which can be attributed to the non-idealities of the bipolar switch. One contribution to pedestal error is the capacitive coupling through the base-collector capacitance $C_{bc}$ of device $Q_2$ during the track to hold transition. This effect is largely canceled in a differential structure, but gives rise to third order distortion which can be quantified by Eq. 3.23 where $V_P$ is the input voltage amplitude, $V_{SW}$ is the clock swing, $V_{ave, out}$ is the DC output voltage, $V_{ave, dr}$ is the average of driving signal, $C_{to}$ is the base col-
lector capacitance, and $\Phi_0$ is a process dependent parameter [22]. Although this error is hard to quantify exactly, the equation provides a basis for minimizing pedestal error.

$$HD_3 = \frac{15 C_{t0}}{32 C_H} \frac{V_{S0}^2 V_{SW}}{(\Phi_0 + V_{ave,\text{out}} - V_{ave,d})}$$

(3.23)

In addition to the distortion contributed by the non-linear capacitance of the differential pair, switching time plays a role in the signal corruption induced by the switching event. Switching in a bipolar SEF occurs during a non-zero period of time, referred to as the aperture time. The aperture time of the SEF can be dictated by the input voltage because of different collector voltages on the switched differential devices. Both legs of the pseudo differential SEF see different voltages at their inputs, therefore their aperture times differ. This variation in aperture time results in asymmetric charging or discharging of the hold capacitors, and inevitably results in distortion, which is commonly referred to as aperture error or aperture distortion [20, 23, 24]. To minimize the difference in aperture time, common practice dictates applying a large signal swing to the switched differential pair. Although this helps equalize the aperture time of the two paths, it increases the switching feedthrough distortion quantified in Eq. 3.23. It’s clear from this argument that selecting a proper clock amplitude is a trade off between aperture error and nonlinear switch feed through.

### 3.5 Hold-Mode Feedthrough

The base emitter junction of $Q_1$ in the SEF has a capacitance which allows a portion of the input signal to feedthrough to the hold capacitor during the hold-phase. This hold-mode feedthrough is a function of the hold capacitance and the parasitic capacitance of the device [20, 25]. Determining the impact of this error is not com-
Figure 3.13: The input signal leaks through the base-emitter capacitance of the switching device.

completely straightforward without knowing more about the conversion time requirement of the quantizer; however, the error can be established for a given feedthrough factor, hold period, and input frequency. The model assumes that the sampling rate is much faster than the input frequency, therefore the slope of the signal is assumed to be linear. The worst case error occurs when the input signal is near the zero crossing [26]. The error at the end of the sampling period is quantified in Eq. 3.24.

\[
Error = \frac{T_{\text{sample}}}{2} \frac{C_{be}}{C_{be} + C_H} A \times 2\pi f_{\text{signal}}
\]  

(3.24)

### 3.6 Sampling Jitter

The resolution of high speed track-and-holds is also affected by nondeterministic aperture uncertainty. This is to be contrasted with the deterministic aperture error which is the result of device non-linearity, not random noise. Sampling jitter is generally attributed to the phase noise of the sampling clock, which results in deviations in the sampling instant, producing amplitude noise. The concept is illustrated for a simple sinusoid in Fig. 3.14. If the sampling time is shifted from a perfectly peri-
odic sampling clock by a time $\Delta t$, the sampled signal incurs an error voltage denoted $\Delta V$. This error voltage is assumed to take on a ‘white-noise’ characteristic which raises the noise floor at the output of the sampling circuit, and dictates the SNR of the system. Because the voltage noise is dictated by the amplitude and frequency of the input signal, sampling jitter places a maximum SNR limitation on the system, as shown in the following derivation.

Figure 3.14: Representation of the amplitude error induced by timing uncertainty.

First, it is assumed that the period of the input signal is much longer than the sampling time error, such that the input signal can be approximated by its derivative around the sampling instant. The voltage error is calculated in Eq. 3.25 for a small timing error around the sampling instant.

$$
\Delta V = \Delta t \times \frac{dV}{dt}
$$

(3.25)

For an input signal $V_{in} = Acos(2\pi f_{in}t)$ the voltage error is expressed in Eq. 3.26.

$$
\Delta V = \Delta t \times 2\pi f_{in} Acos(2\pi f_{in}t)
$$

(3.26)
To generate an expression for maximum SNR, the RMS voltage error must first be determined. This determination is made by first assuming that the timing error has a Gaussian distribution, with an RMS value of $\Delta t_{RMS}$.

\[
\Delta V_{RMS} = \Delta t_{RMS} \times \sqrt{\frac{1}{T} \int_0^T (2\pi f_{in} A \cos(2\pi f_{in} t))^2 \, dt}.
\]

The RMS voltage error is simply the product of the RMS timing error and the RMS value of the derivative as shown in Eq. 3.27. This ‘white noise model’ allows for a simple quantification of error, results in a maximum SNR for a given RMS timing error and input frequency [20]. If no other noise is present in the system, its SNR is limited to the value determined by Eq. 3.29. If noise is introduced by other mechanisms, the SNR of the system can only be further degraded. The result of this derivation is an upper bound on system resolution when sampling jitter is present.

\[
SNR_{max} = \frac{\left(\frac{A}{\sqrt{2}}\right)^2}{(\Delta t_{RMS} \times \sqrt{2\pi f_{in} A})^2} \quad (3.28)
\]

\[
SNR_{max}(dB) = 20 \log \left( \frac{1}{\Delta t_{RMS} \times 2\pi f_{in}} \right) \quad (3.29)
\]

Circuit techniques can be used to eliminate or reduce most of the error sources mentioned above. Specific implementations will be discussed in the design portion of the following sections of this work.
Chapter 4

A 40GS/s High Resolution Track-and-Hold

The following chapter introduces a 40GS/s track-and-hold topology designed to overcome some of the non-idealities of the differential SEF architecture introduced in Chapter 3. This work was done in the 250nm InP on CMOS process introduced in Chapter 1. The design was intended to perform with greater than 8 ENOB at an input frequency of 1GHz, with power consumption comparable to existing high resolution architectures. Section 4.1 discusses common methodologies to improve sampling resolution in open-loop SEF architectures. Section 4.2 discusses the design of the proposed high resolution track-and-hold and Section 4.3 covers the design of the high linearity output buffer used to drive the waveform off chip. Finally, Section 4.4 discusses the performance of the complete system and compares these results with other published works.
4.1 High Resolution Compensation Techniques

The architecture introduced in this section is designed to reduce the affect of hold-mode feedthrough and pedestal error inherent to SEF track-and-holds. Most high-resolution track-and-hold circuits employ techniques to minimize the effect of hold-mode feedthrough discussed in Section 3.5. Common techniques used to minimize this error are 1) use a pair of cross coupled capacitors to cancel the parasitic effects of the base-emitter junction of the emitter follower or 2) use a switched cross coupled buffer to regulate the input voltage during the hold cycle.

The first approach is purely a passive attempt at correcting the error. This approach is highlighted in Fig. 4.1 where two capacitors $C_p$ are cross coupled from input to output to cancel the input signal feedthrough. $C_p$ is sized to mimic the parasitic capacitance of each switch in the differential SEF. This approach is not as easy to implement as it might appear. In general the feedthrough attenuation capacitors cannot be simple metal/dielectric devices because their process dependent variation does not track the capacitance of the base-emitter junction, which by itself...
is also hard to predict due to its bias dependence. The common solution for this is to implement $C_p$ as a series-parallel network of bipolar devices as shown in the inset of Fig. 4.1. This configuration reproduces the base emitter junction capacitance of the switch in a process invariant manner.

The second approach is to use a cross-coupled buffer at the input to the track-and-hold which is switched complementary to the SEF. During the hold period the secondary buffer applies the differential signal to the input of the SEF, with the signal polarity reversed. Ideally this approach cancels the time varying portion of the input swing, forcing the input to remain constant during the hold phase. Without amplitude variation the input signal does not leak through the parasitic capacitance of the switch, and therefore does not distort the output. A simple diagram of this architecture is shown in Fig. 4.2. In literature the secondary buffer is commonly referred to as a, ‘feedthrough attenuation’ buffer [27, 28].

The first approach is fairly simple to implement and requires no DC power from the supply, but it falls short when device mismatch is considerable. For this reason, the cross coupled architecture has an upper limit to its effectiveness, dictated by process non-uniformities. The feedthrough attenuation architecture eliminates the need for precise device matching by simply eliminating the input signal. This comes
at the cost of added DC power consumption and certain negative side effects which aren’t immediately apparent and are discussed in the following section.

4.2 40GS/s Double-Switching Track-and-Hold

The feedthrough attenuation architecture pulls the differential input nodes to the same voltage during the track-to-hold transition. A common input potential forces an imbalanced in switching time, worsening pedestal error and degrading the performance of the circuit. The proposed architecture mitigates this negative outcome such that both feedthrough attenuation and pedestal error reduction can be implemented simultaneously. The proposed system uses a double switching-architecture as previously discussed in [29]. The novelty to this approach is to leverage the speed of the InP devices in order to apply these corrective techniques to sampling systems operating at 10’s of GHz. The proposed architecture is shown in Fig. 4.3. Aside from the input buffer, clock driver, SEF and output buffer, the schematic shows a number of blocks which help to make up the double-switching architecture. The block labeled

![Diagram of proposed double-switching architecture](image)

Figure 4.3: Proposed ‘double-switching’ architecture.
‘SB’ is a switching buffer which switches the SEF input signal between the buffered input signal and feedback signal. During the ‘track’ phase the switched buffer applies the buffered input signal to the input of the SEF, which passes the signal onto the hold capacitor. During the track-to-hold transition the switched buffer applies the feedback signal to the input of the SEF as it switches off. During the ‘hold’ phase, the SEF sees a constant voltage at its input which provides the strong feedthrough isolation desired in high resolution applications. In addition to the improvement in hold-mode isolation, the feedback buffer plays a critical role in minimizing pedestal error. This improvement comes from switch feedthrough and delay equalization in the differential paths of the SEF.

In the case of the feedthrough attenuation architecture, the input voltages on the differential SEFs are pulled to the same DC level during the track-to-hold transition. This is shown in Fig. 4.4a. This action produces unequal base-emitter potentials on the two switching devices, due to the differential output voltages held on the capacitors. This produces additional undesirable pedestal error, which the proposed architecture seeks to resolve. In the case of the double switching architecture the feedback buffer applies the output signal to the input, rendering the differential input voltage the same as the differential output voltage; therefore, the switching potential across the junctions of the emitter followers becomes a common mode transition. This concept is illustrated in Fig. 4.4b.

The common mode transition helps minimize pedestal error in two ways. First, the junctions of the two emitter followers see a common mode transition, which ideally equalizes their track-to-hold transition time. Equalizing this transition time prevents skew in the signal leaked through each switch onto the hold capacitors during the transition. Secondly, with the feedback buffer in place the input to each SEF transitions with the same voltage slope as illustrated in Fig. 4.4b. As in the case
Figure 4.4: Illustrates the advantage of the double switching architecture. (a) Result of the conventional architecture. (b) The double switching architecture provides an equal voltage drop ($\Delta V_1 = \Delta V_2$) at the input of both switches.

of input feedthrough, the input switching transition can leak through the parasitic capacitance of the junction and distort the voltage on the hold capacitor, which is an additional source of pedestal error induced by the conventional feedthrough attenuation architecture. The common mode transition of the double switching architecture is designed to overcome this. The transition signal feeds through the parasitic capacitances of the device junctions, but in common mode. Thus the distortion is induced on the hold capacitors as a common mode error, and rejected at the differential outputs.

The double-switching buffer is shown integrated with the feedback buffer in Fig. 4.5. These blocks produce the desired affect described earlier. During the ‘track’ period, the input signal is applied to the load resistors $R_L$, and fed to the input of the SEF. The feedback buffer applies the output signal to the dummy resistors labeled $R_{dummy}$. During the hold phase, the input signal is applied to the dummy resistors, and the output signal is applied to the load resistors and fed into the SEF. Driving the input signal across the dummy resistors during the hold phase produce
a high degree of hold mode isolation, alleviating the issue of hold-mode feedthrough. Given the conditions above, the double-switching architecture helps to alleviate hold-mode feedthrough without incurring additional pedestal errors inherent in other active feedthrough attenuation techniques.

Additional specifics regarding the design of the track-and-hold can be found in [30]. The track-and-hold, clock driver, double-switched buffer and feedback buffer consume a total of 51mA from a 6V supply making the power consumption of the sampling system roughly 300mW.
4.3 Output 50Ω Buffer

Given the resolution of the high speed track-and-hold, a buffer was required to drive the high resolution signal off-chip without degrading its linearity. A cascoded resistively degenerated differential pair was adopted to fit the requirements. The cascode topology was used to minimize the loading on the transconductor and improve its linearity. Degeneration was used in the form of a 30Ω resistor at the emitters of each driving device. In order to support the power consumption of the buffer, 6 parallel devices were used in each leg of the differential pair. This included 6 driving devices and six cascode devices. At each output terminal of the buffer there is effectively a 25Ω load, which is a parallel contribution of 50 ohms seen at the RF probes and another 50 ohms from the termination resistor. The gain of the buffer is slightly lower than unity due to the degeneration resistance being set higher than the

![Schematic of the linear output buffer.](image)

Figure 4.6: Schematic of the linear output buffer.
load. The gain of the buffer is calculated in Eq. 4.1.

\[ A_v \approx \frac{R_{load}}{R_{degen}} = \frac{25\Omega}{30\Omega} = -1.6dB \]  

4.1

First, simulation results are presented which demonstrate that the buffer was designed with adequate bandwidth. The simulated bandwidth of the buffer is in excess of 70GHz when driven with a 25Ω source impedance which is a fair approximation of the track-and-hold output impedance. A slight reduction in the low frequency gain can be noted in the simulated result Fig. 4.7 which is due to the approximation made in Eq. 4.1.

Figure 4.7: The simulated small signal gain and S\text{22} parameter of the 50Ω output buffer.
Next it was important to consider how well the buffer was matched to its off-chip environment. The return loss was simulated from DC-80GHz and demonstrated a -10dB bandwidth of 28.3GHz. Although the buffer should ideally be matched above 40GHz, it was concluded that the result was sufficient to mitigate any significant reflection during measurement. Once the losses of the cabling and input return loss of the measurement equipment is taken into account the simulated return loss of the buffer appears sufficient. The results for the s-parameter simulation are shown alongside the AC gain results in Fig. 4.7.

The buffer was designed for high linearity and as a metric to prove this, the IM$_3$ of the buffer was simulated from 1-50GHz. Fig. 4.8 shows the distortion produced by the buffer. For an input two tone test with 600mVpp-differential envelope, the simulated IM$_3$ of the buffer is lower than -64dBc over this entire frequency range, and is lower than -77dBc at 1GHz which is greater than 12.5 ENOB. The buffer and bias networks draw 66mA from a 6V supply, which results in a total power consumption close to 400mW.
4.4 Performance of the 40GS/s Track-and-Hold

Simulation confirms the performance of the high-resolution track-and-hold and output buffer. Fig. 4.9 shows the output spectrum of the system for a two tone test, at 1 and 1.1 GHz respectively. The envelope of the simulated input signal was 600mVpp differential. The output $IM_3$ product was -76.7 dBc, corresponding to greater that 12 ENOB. A time domain waveform is also included in the figure.

The simulated performance of the high resolution track-and-hold is compared with other compensated architectures in Table 4.1. This work shows a significant improvement over [28] in area, power consumption, sample rate and harmonic distortion, but consumes more power than [27] although achieving higher linearity. Both of the

Figure 4.9: Output waveform and spectrum of the high-resolution track-and-hold. $F_{in} = 1GHz/1.1GHz$ and $F_s = 40GS/s$. 
works being compared to this work use the feedthrough attenuation architecture described in Section 4.1.

Table 4.1: Compensated Track-and-Hold Performance Comparison

<table>
<thead>
<tr>
<th>$F_s$</th>
<th>Linearity Amplitude @ $F_{in}, F_s$</th>
<th>Supply (V)</th>
<th>Power</th>
<th>Process</th>
<th>Area (mm$^2$)</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>40GS/s</td>
<td>-76.7dBc IM$_3$ 600mVpp-differential @ 1GHz, 40GS/s</td>
<td>+6</td>
<td>690mW</td>
<td>InP</td>
<td>1.1 x 0.7</td>
<td>This Work</td>
</tr>
<tr>
<td>40GS/s</td>
<td>-32.4dB THD 1Vpp-differential @ 10GHz, 40GS/s</td>
<td>+5.5</td>
<td>560mW</td>
<td>SiGe</td>
<td>Core 0.8 x 0.2</td>
<td>[27]</td>
</tr>
<tr>
<td>20GS/s</td>
<td>-38dB THD 500mVpp-differential @ 0.9GHz, 20GS/s</td>
<td>-5.2</td>
<td>735mW</td>
<td>InP</td>
<td>2.0 x 2.0</td>
<td>[28]</td>
</tr>
</tbody>
</table>

The track-and-hold introduced in this chapter is useful in broad band high-resolution applications like tests and measurement or optical sampling. The high linearity of the track-and-hold enables high resolution sampling with greater than 12 ENOB and comparable power consumption with other lower linearity systems. This system is an advancement on state of the art, broadband, high sample rate data converters.

4.4.1 Chip Layout

The layout of the complete system is shown in Fig. 4.10. Input and output pins are on the left and right of the chip respectively. The clock is fed from the bottom side of the chip. The track-and-hold and clock buffer occupy the left side of the chip. The chip measures 1,062µm by 712µm. Ground metal fill has been removed from the image for clarity.
Figure 4.10: Layout of the high resolution track-and-hold and output buffer.

**Acknowledgments**

Chapter 4 discusses a project which was a collaboration between the Thesis author and Timothy D. Gathman. Design of the double switched track-and-hold was completed by Dr. Gathman, and is included in his Doctoral Dissertation [30].
Chapter 5

A Time Interleaved 40GS/s BiCMOS Sampling System

Chapter 5 discusses the design of a 40GS/s time interleaved system fabricated in the InP/Si BiCMOS process discussed in Chapter 1. This time interleaved system was developed as an extension to the work presented in Chapter 4, with the purpose of demonstrating enhancements in system performance when implemented in a BiCMOS technology while, bringing the entire system closer to a complete data converter. Data converters supporting the rates and resolutions necessary for tests and measurement equipment and optical sampling often require time interleaving architectures, which enable operating speeds that single channel data converters are incapable of attaining.

The InP HBT devices posses superior high frequency performance and were thus used at the front end of the system. The CMOS devices allow for higher levels of integration which make them more fit to perform in the back end, where multiple channels are run in parallel. Their linearity and low power consumption enhances their usefulness in a multi-channel system, in addition to enabling background calibration in both the bipolar and CMOS blocks. It is for this reason that a BiCMOS
6.8 process is highly useful, if not required, in such a demanding application.

5.1 Time Interleaved System

A highly linear time interleaved sampling system is proposed in this section, for use in a high-resolution high-speed data converter. The time interleaved system consists of a highly linear buffer, clock driver, divider and 8 parallel sampling circuits. The architecture is shown in Fig. 5.1. Each individual interleave was designed to operate at a rate of 5GS/s. The individual building blocks of the system are shown in the following sections.

Figure 5.1: Time interleaved architecture, consisting of 8 interleaves.
5.2 High Linearity Wideband HBT Buffer

When designing the time interleaved system it became apparent that a large buffer was necessary to drive all 8 channels of the interleaved CMOS system. This required high linearity and wide bandwidth with modest power consumption. Initially a simply emitter follower was considered because of its high linearity and low output impedance; however, in order to drive the capacitance of the 8 interleaved cells the power consumption of the emitter follower was unacceptably high. It was with this in mind that the adapted architecture was used. This modified architecture is shown in Fig. 5.2 where Q₃ mirrors current to the output of the buffer, effectively boosting the gm of the emitter follower. This gm-boosting improves the linearity and bandwidth of the buffer simultaneously with minimum power addition. Devices Q₄ and Q₅ are placed in the feedback loop to reduce the voltage swing across Q₃.

![Diagram of Wideband HBT buffer](image)

Figure 5.2: Wideband HBT buffer using gm-boosting.

The gm-boosting in the buffer can be modeled by representing the current mirror as a linear voltage controlled current source, with a finite input impedance. If the buffer is loaded by an impedance $Z_L$, the current through the load is found in
Eq. 5.1. The non-linear transconductor produces a voltage across $Z_F$ which in turn controls the current produced by the current mirror. This voltage is shown in Eq. 5.2 and is opposite in sign to the output voltage. Therefore the added current is in phase with the current from the emitter follower which increases the effective gm of the buffer. Each of the transconductances terms take on a new effective transconductance which is $1 + gm_F Z_F$ times larger than the original terms Eq. 5.3.

$$i_{out} = gm_1 v_{be} + gm_2 v_{be}^2 + gm_3 v_{be}^3 - gm_F v_F$$
$$v_f = -(gm_1 v_{be} + gm_2 v_{be}^2 + gm_3 v_{be}^3) Z_F$$
$$i_{out} = gm_1 (1 + gm_F Z_F) v_{be} + gm_2 (1 + gm_F Z_F) v_{be}^2 + gm_3 (1 + gm_F Z_F) v_{be}^3$$

This boost in the gm of the buffer improves both the linearity and bandwidth of the buffer. To first order the improvement in linearity can be explained by considering that as the gm of the buffer increases the better the output voltage tracks the input. If the output tracks the input more closely there is less room for the error generated by non-linearities, hence a more linear response. This intuition is confirmed by the results of the Volterra analysis, where the gm of the device is replaced by the effective ‘gm’ of the buffer.

Eq. 5.4 captures the theoretical result for $HD_3$ from the Volterra analysis conducted in Section 3.2. As the analysis will show, gm boosting decreases $HD_3$ for
the modified buffer.

\[ HD_3 \approx \frac{A^2}{4} \left| \frac{H_3(j\omega_o, j\omega_o, j\omega_o)}{H_1(j\omega_o)} \right| \]  

(5.4)

For highly degenerated emitter followers (i.e. \( Z_1G_1 \gg 1 \)) the first and third order Volterra operators can be simplified as shown in Eq. 5.5 and Eq. 5.6 respectively.

\[ H_1(\omega_o) = \frac{G_1Z_1(\omega_o)}{1 + G_1Z_1(\omega_o)} \approx 1 \]  

(5.5)

\[ H_3(\omega_o, \omega_o, \omega_o) = \frac{Z_1(3\omega_o)}{(1 + G_1Z_1(\omega_o))^3(1 + G_1Z_1(3\omega_o))} \left( G_3 - \frac{2G_2^2Z_1(2\omega_o)}{(1 + G_1Z_1(2\omega_o))} \right) \]

\[ \approx \frac{Z_1(3\omega_o)(G_3G_1 - 2G_2^2)}{G_1(G_1Z_1(\omega_o))^3(G_1Z_1(3\omega_o))} \]  

(5.6)

The third harmonic distortion can then be approximated by Eq. 5.7.

\[ HD_3 \approx \frac{A^2}{4} \left| \frac{Z_1(3\omega_o)(G_3G_1 - 2G_2^2)}{G_1(G_1Z_1(\omega_o))^3(G_1Z_1(3\omega_o))} \right| \]  

(5.7)

Each of the transconductance terms are increased by the feedback, which results in a new expression for third order distortion which is denoted \( HD'_3 \). If each transconductance term is increases proportionally by a factor \((1+X)\), the expression for \( HD'_3 \) is approximated by Eq. 5.8

\[ HD'_3 \approx \frac{A^2}{4} \left| \frac{Z_1(3\omega_o)(G_3(1+X)G_1(1+X) - 2G_2^2(1+X)^2)}{G_1(1+X)(G_1(1+X)Z_1(\omega_o))^3(G_1(1+X)Z_1(3\omega_o))} \right| \]

\[ \approx \frac{HD_3}{(1+X)^3} \]  

(5.8)

As shown in the analysis earlier in this section, gm boosting effectively increases the three transconductance terms of the emitter follower by a factor \( 1+gm_FZ_F \).
result is an improvement in third order distortion captured in Eq. 5.9.

\[
HD_3' \approx \frac{HD_3}{(1 + gm_F Z_F)^3}
\]  

(5.9)

Bandwidth is improved in the modified buffer because the feedback loop reduces the output impedance of the buffer. The dominant pole of the buffer is caused by the output impedance of the buffer, and the large load capacitance. As the output impedance is reduced the dominant RC pole is pushed out in frequency, broadening the bandwidth of the circuit.

Several simulations confirm the intuition gained from the analysis of the small signal model. The modified buffer in Fig. 5.2 was simulated against a simple emitter follower. Both circuits consumed the same amount of power and drove the same 400fF (single-ended) load. The load is a fair representation of the loading caused by the time interleaved CMOS circuits. Fig. 5.4 illustrates the results of a PSS simulation which was run over frequency, and captures the IM₃ performance of the circuit. The input to each buffer was held constant at 600mVpp differential. The modified buffer is more linear than the emitter follower up to about 14GHz, where effects not considered in the simple small signal model degrade the linearity performance of the circuit. This is likely due to delay in the feedback loop. At low frequency the modified buffer shows IM₃ approximately 15dB below the simple emitter follower. This design uses a gm enhancement \((1+gm_F Z_F) \approx 3.24\), which results in a theoretical reduction in third order distortion of 15.3dB. This matches the simulated result very closely, confirming the outcome of the theoretical analysis.

Fig. 5.5 shows the bandwidth improvement in the modified architecture. The -3dB bandwidths of the modified and simple emitter follower were 30.9GHz and 16.1GHz respectively. The modified architecture has nearly twice as much bandwidth as the simple emitter follower. It can be seen from this AC simulation that
there is quite a bit of peaking in the response of the modified emitter follower. Stability becomes a concern in a design such as this. Adequate steps were taken to ensure that the buffer was stable with approximately $90^\circ$ of phase margin.

Figure 5.5: The small signal AC response of the modified and simple emitter follower.
5.3 CMOS Sampling System

The architecture of each of the CMOS sampling circuits is shown in Fig. 5.6. The input signal is sampled using a composite NMOS switched and held on the two capacitors labeled $C_H$. The signal is then buffered to the next switching stage by a highly linear buffer. The buffer is further linearized and bandwidth extended with a gm-boosting technique, discussed in Section 5.3.2. The next switching stage consists of a compound PMOS switch which is used to track-and-hold the previously sampled signal. After the PMOS switching stage the signal is buffered by a 50Ω buffer to the output pads of the chip. The output buffer is discussed in Section 5.3.3.

![Figure 5.6: The CMOS sample-and-hold architecture.](image)

5.3.1 NMOS & PMOS Track-and-Hold

Each track-and-hold consists of differential switches, feedthrough cancellation devices, and hold capacitors. The switches were sized to meet the bandwidth requirements of the system. The PMOS switches were sized larger because they have higher ‘on’ resistance than the NMOS devices. Aside from this difference the N and P-MOS track-and-hold circuits are identical. A simple diagram of the architecture is shown in Fig. 5.7. The block denoted ‘F’ represents the feedthrough cancellation device.
Using a MOS device as a switch presents some nonidealities which must be addressed to achieve high resolution sampling. The first issue arises when the switching device is in its on state. When the device is on, charge is stored in the channel of the device which in the case of an NMOS, is the result of electrons being pulled to the interface of the oxide and the substrate when a positive voltage is applied. When the voltage is removed and the potential on the gate reduces, the charge is released. Fig 5.8 represents this action. This has a profound effect on the resolution of MOS sampling switches, because some of that charge is dumped on the hold capacitor and corrupts the held signal voltage. If the switches are used differentially this effect can be somewhat neglected as a common mode contribution; however, odd order distortion arises from the nonlinearity of this charge accumulation.

\[ Q_{\text{ch}} = -WLC_{\text{ox}}(V_{GS} - V_T) \]  

(5.10)

When the device is on and the channel is active, and there is a charge which can be quantified by Eq. 5.10 after [31, 32]. This stored charge is a non-linear function of the input voltage, and is therefore a source of distortion. A common technique to resolve this problem is to place two half sized dummy transistors on opposing sides
of the switching device. These dummy transistors are then clocked with a signal complementary to the switch. The desired affect is a reabsorption of the injected charge by the two dummy devices [33]. Ideally this action negates the effect of charge injection and prevents signal corruption at both the input and output of the track-and-hold. Theoretically this method of cancellation is limited only by mismatch.

Figure 5.9: Representation of the redistribution of charge. (a) Channel charged with $Q_{ch}$ when the switch is on. (b) Charge redistributes to the dummy devices when the switch is turned off.

Note that the two dummy devices are shown with half the length of the switch-
ing device. This unconventional representation was used only to simplify the diagram. Generally designers using this technique halve the width of the dummy devices to compensate charge. The present design uses half width devices for compensation.

The second nonideality which degrades the performance of the MOS switch is its non-zero off capacitance. When the switch is off there is a small amount of capacitance which allows the input signal to leak through the switch and distort the held voltage. This capacitance is primarily due to the gate overlap capacitance \( C_{ov} \) but can also be caused by substrate coupling. Fig. 5.10 represents the ‘off’ state model of the switch. \( R_{Driver} \) models the impedance at the output of the switch driver, which forms a high pass filter with the gate overlap capacitances. \( R_{sub} \) models the effective substrate resistance between the source and drain diffusions. If this resistance is large enough the substrate coupling can be ignored, and the gate coupling becomes the primary contributor to hold-mode distortion.

A common solution to this design problem is to cross couple two devices - which are the same size as the switches - across complementary input and output nodes. This method provides a complementary path for signal feedthrough which allows for cancellation at each of the outputs. The driver output impedance can be
replicated by placing a resistance from the gate of the feedthrough device to ground.

Additional signal distortion arise from the nonlinearity of the switch itself. The primary contribution to harmonic distortion during the ‘track’ phase, is the nonlinear ‘on’ resistance of the switch. As Equation 5.11 shows, the ‘on’ resistance of the device is a function of the voltage \( V_{GS} \) [21]. This input dependent resistance causes a nonlinearity, because the time constant dictated by \( R_{on}C_H \) is a nonlinear function of the input voltage. This is a source of dynamic nonlinearity that is sometimes referred to as aperture delay distortion.

\[
R_{on} = \frac{1}{\mu C_{ox} \frac{W}{L} (V_{GS} - V_{TH})} \tag{5.11}
\]

To eliminate the contribution of this non-linearity, the switch should be sized and driven to minimize its ‘on’ resistance. As the ‘on’ resistance of the switch decreases, less voltage is dropped across its terminals and the non-linear modulation is reduced. The device can be sized very large to improve the resistance, but this method comes at the cost of additional parasitic capacitance, which can harm bandwidth and itself cause additional nonlinearities. The other approach is to drive the switch with a high gate potential. This method was favored, which lead to the design of CMOS level switch drivers for this system. The CMOS level switching brings the gate potential near the breakdown of the device which minimizes the device ‘on’ resistance and provides a simple solution to driving the gates.

The complete design of the N and P-MOS track-and-hold blocks are shown in Fig. 5.11 and Fig. 5.12 respectively. Each track-and-hold includes devices to counteract the affect of charge injection, and signal feedthrough. The hold capacitors were sized to be \( 40\text{fF} \) and implemented with MIM caps for high linearity.
Figure 5.11: The NMOS track-and-hold.

Figure 5.12: The PMOS track-and-hold.
5.3.2 gm-Boosted Buffer

This buffer was a necessary stage in the CMOS sampling circuit because it prevents the second track-and-hold stage from draining charge from the hold capacitor during the hold stage of the first track-and-hold. Like the other stages in the sampling chain linearity was optimized to ensure that this stage was not a bottleneck for harmonic distortion. The buffer used in the sample-and-hold utilized a gm-boosting technique similar to the wideband HBT buffer. Current is injected by the PMOS current mirror, which effectively increases the gm of the buffer. This improves linearity and bandwidth at the cost of slightly increased power consumption. The design was adapted from a buffer used in [34].

![Figure 5.13: The gm-boosted buffer.](image)

Simulations were run on this buffer while the PMOS track-and-hold was in track mode. Putting the proceeding stage in track mode represents the maximal loading on the buffer and therefore the worst case linearity. The output 50Ω buffer
was also included in the simulation, because it contributes to the loading on the PMOS buffer. The bandwidth of the buffer was also measured with the loading mentioned above. The -3dB bandwidth of the buffer is 7.4GHz and its low frequency gain is approximately -2.3dB. The results of this simulation are shown in Fig. 5.14. This loss is due to the body effect in the source follower. There are methods of compensating for this loss, one of which consists of driving the body of the device with a parallel source follower. This can only be done in triple well, or floating body technologies and comes at the expense of bandwidth and added power consumption.

The linearity of the buffer was measured using a two tone test and was swept from 1-30GHz. The envelope of the input signal was 600mVpp differential. This represents the worst case amplitude used to drive the buffer. The results of this simulation are shown in Fig. 5.15. The simulation suggests that the buffer is capable of providing better than -67dBc IM3 at 1GHz, and better than -35dBc over the entire frequency range. Each buffer draws 2.55mA from a 1.2V supply. The two pseudo-differential buffers draw 6mW from the same 1.2V supply.

Figure 5.14: AC small signal bandwidth of the PMOS buffer.
5.3.3 NMOS 50Ω Output Buffer

A high linearity output buffer was designed to drive the sampled signal off chip and match to the 50Ω measurement environment. The three main design criterion for the buffer was that it had adequate bandwidth and linearity and that it was matched to the off chip environment over an appropriate bandwidth. A pseudo-differential source follower architecture was used in the design of this buffer. The output is

Figure 5.15: IM₃ distortion in the PMOS gm-boosted buffer from 1-30GHz.

Figure 5.16: 50Ω pseudo-differential output buffer.
further degenerated by the series 25Ω resistor. The resistor helps to further linearize the circuit as well as provide output matching but comes at the cost of gain.

![Graph showing IM3 distortion in the NMOS output buffer from 1-30GHz.]

Figure 5.17: IM3 distortion in the NMOS output buffer from 1-30GHz.

![Graph showing AC gain and S22 parameter of the buffer.]

Figure 5.18: The simulated AC gain and S22 parameter of the buffer.

In order to confirm that the buffer was adequately linear the IM3 of the buffer was simulated from 1-30GHz and is shown in Fig. 5.17. There is almost no degradation
in linearity over this band. IM₃ remains below -55dBc equivalent to greater than 8 ENOB. The small signal AC bandwidth of the buffer was simulated and is shown in Fig. 5.18.

5.4 8 Phase Clock Generator

In order to drive the 8 time interleaved channels of the sampling circuit, an 8 phase clock generator was designed. In this implementation a divider tree was utilized. A single differential divide-by-two stage produces two signals which are shifted by 90° relative to one another. If the polarity of the differential signals is flipped two more phases are available. If the 0° and 90° phase of the first divide-by-two stage is cascaded into parallel secondary divide-by-two stages, 8 available phases are produced. If the positive polarity is considered there are four phases which are produced at 0°, 45°, 90°,
and 135° relative to the input phase. Using the opposite polarity generates phases at 180°, 225°, 270°, and 315° relative to the input. This architecture does not produce deterministic phases, but because only one output was capable of being measured, the ordering of the phases was determined to be irrelevant. In a complete time interleaved data converter phase determinism is a necessary requirement, because if the sampled signals are not properly phase shifted the reconstructed signal will be meaningless.

Fig. 5.19 shows a schematic illustration of the 8 phase generator. Between the first divider and the parallel secondary dividers, there are two CML buffers. One for each of the I and Q paths (they are referred to in this way for simplicity, because of their 90° phase relation). These buffers were required to boost the amplitude of the divided signal into the next divide-by-two stages.

### 5.4.1 Divide By 2

Each divide-by-2 block was produced using a pair of cross coupled CML latches shown in Fig. 5.20. The CML latches were then clocked with complementary input signals generating an edge triggered flip-flop. The polarity of the output signal is flipped and fed back into the input. The effect is a signal which changes state at the rate of the clock edge and therefore has half the frequency of the clock. Each

![Figure 5.20: The divide by 2 circuit.](image)
divide-by-two circuit generates a differential I and Q phase relative to the input clock phase.

Fig. 5.21 represents the architecture used to generate the latching action. When the clock signal is high, the input signal is buffered to the output. When the clock transitions from high to low, the secondary differential pair amplifies the output signal. The positive feedback in the secondary amplifier locks the output during the low phase of the clock. Therefore the output tracks the input during the high phase and is held during the low phase, hence the latching action is produced. Each divide-by-two stage draws an average of 8mA from a 1.2V supply, consuming roughly 9.6mW.

### 5.4.2 CML Buffer

The CML buffer was designed to boost the signal level generated by the divide-by-two stage, in order to drive the input of the following stage. The current in the buffer is switched from one leg to the other depending on the polarity of the input.
This ideal switching allows for a quick determination of the output signal swing. The current through the tail current source of each buffer is 4mA, which generates a 1.6V$_{pp}$ differential output. This swing is diminished to roughly 1.1V$_{pp}$ at high frequency because of additional loading at the output nodes of the buffer. This swing is sufficient to drive the input of the following divide-by-two stage. The power consumption of each CML buffer is 4.8mW from a 1.2V supply.

![Figure 5.22: Differential CML buffer.](image)

### 5.4.3 CML To CMOS Converter

The CML to CMOS converter consists of a pair of differential to single ended converters, which are connected complementary to one another. Each converter takes the low swing differential CML signal and produces a high swing output signal. The high swing signal is necessary to drive the following CMOS inverter chain. Each pair of differential-to-single-ended converters draws an average of 3mA from a 1.2V supply, consuming 3.6mW.
Fig. 5.23: A CML to CMOS converter.

Fig. 5.24 shows a post extraction simulation of the 8 phase clock generator.

Figure 5.24: The 8 phases of the clock are generated from a 20GHz input.
The input waveform is a 20GHz sinusoid. Transient simulations were run to confirm that the extracted model could operate above the design requirement. The 8 phase generator was capable of operating beyond 23GHz, which allowed for 15% margin in the design.

5.5 A Cascaded Track-and-Hold & Time Interleaved System

As discussed in Section 2.3 circuit techniques can be used to minimize timing errors inherent in time interleaved systems. It was with this in mind that the cascaded track-and-hold / time interleaved system was designed. The architecture in Fig. 5.25 was designed in the InP/Si BiCMOS technology, using the track-and-hold discussed in Chapter 4 and the time interleaved system discussed in this chapter. There are a

![Diagram](image-url)
number of advantages to cascading the two systems, which are primarily due to the
hold pedestal produced by the input track-and-hold.

First, the hold pedestal - if aligned properly in time - produces a zero derivative signal at the instant any of the time interleaved sample-and-holds capture the signal. As mentioned in Chapter 2, timing skew between channels places a frequency dependent SFDR limitation on the interleaved system. If the input track-and-hold is in hold mode, the signal appears to be DC, therefore clock skew between channels has no affect on the SFDR of the system. This is true, provided that the clock skew is limited to within the hold phase of track-and-hold. The second benefit is that the hold phase allows for additional settling time in the interleaved channels, which decreases the bandwidth requirements of the time interleaved channels [35].

In addition to the correcting systematic timing error, the cascaded architecture helps improve clock jitter. Intrinsically the HBTs impart less timing jitter on their sampling clock, than the CMOS devices in the back end of the system. In the proposed architecture, the CMOS switches sample the signal on the hold phase of the input track-and-hold. This zero derivative signal does not deviate in time, thus the jitter in the CMOS sampling clock does not result in an additional amplitude error. In this case the jitter error of the system is dictated by the aperture uncertainty of the HBT track-and-hold, not the CMOS sampling circuits. The error induced by the HBTs is ideally much lower than the CMOS devices, improving jitter in the system [36].

5.6 System Performance

This section includes simulation results for the system described in this chapter. First the results of the time interleaved CMOS sampling system are presented.
A time domain waveform and frequency spectrum are shown in Fig. 5.26. The plots are shown for a two tone input signal at 1 and 1.1GHz respectively. The amplitude of each signal is 300mVpp-differential, resulting in an input envelope of 600mVpp-differential. The spectrum of a two tone input is shown for an input amplitude of 600mVpp, which was the design goal. The IM$_3$ product in this simulation is -61.6 dBc, resulting 9.9 ENOB. The linearity meets the design requirement of 8 bits, with a fair amount of margin. The time interleaved system draws 12.4mA from a 6V supply, and 98mA from a 1.2V supply. The total power consumption of the system is approximately 190mW. When cascaded with the high-resolution track-and-hold the system draws close to 500mW.
This architecture presents a highly linear approach to sampling in BiCMOS processes, that can be applied to time interleaved data conversion systems. Various system blocks benefit from being designed using bipolar devices, while others benefit from being designed in CMOS. This work depicts some of these advantages in a step towards a fully integrated time interleaved data converter for high-speed high-resolution applications. The work presented is useful in state of the art optical sampling and test and measurement applications.

5.6.1 Chip Layout

The layout of the time interleaved system is shown in Fig. 5.27. The input is applied at the left of the circuit where it is fed into the high linearity HBT buffer. The clock tree is fed from the left of the chip and the output is taken from one of

Figure 5.27: Layout of the time interleaved MOS sampling system.
the interleaved channels which runs to the bottom left side of the chip. The chip measures 810µm by 785µm. The ground metal fill was removed from the image for clarity. Also included is a layout of the high-resolution track-and-hold described in Chapter 4, feeding the time interleaved CMOS system. This chip measures 1,310µm by 910µm.

Figure 5.28: Layout of the high-resolution track-and-hold and time interleaved CMOS sampling system.

Acknowledgments

Chapter 5 discusses a project which was a collaboration between the Thesis author and Timothy D. Gathman. Dr. Gathman has included this work in his Doctoral Dissertation [30].
Chapter 6

An HBT Track-and-Hold Using a Passive ‘Transconductor’

In this chapter a novel track-and-hold concept is presented which uses a passive ‘transconductor’ to replace the active devices used in most applications. A transconductor is generally used to drive a bipolar track-and-hold, because it presents high output impedance to the SEF, which is required to switch the emitter follower on and off. This block is generally implemented with active devices, which suffer from non-linearities. This chapter discusses replacing the active devices with a passive balun or transformer possessing high common-mode impedance, therefore acting as passive common-mode ‘transconductors.’ The term, ‘passive transconductor’ will be used loosely in the following sections to refer to these circuit elements.

Section 6.1 presents a generic implementation of an active transconductor in an SEF track-and-hold and discusses the use of a passive device in its place. Section 6.2 presents a single ended to differential track-and-hold using a balun at its input and Section 6.3 discusses a design using a transformer. Sections 6.4 & 6.5 discuss the output buffer and clock driver used in both systems, and Section 6.6 presents system performance, concluding comments and chip layouts.
6.1 Replacing the Active Transconductor with a Balun

A balun possesses a number of characteristics which makes it an interesting alternative to the active transconductor which is typically used to drive a differential SEF track-and-hold. The classic design of an SEF track-and-hold typically uses a differential common emitter stage which provides enough output impedance to the differential SEF in order for the switching action to occur. In high resolution applications where linearity is critical, the common emitter stage tends to become a bottleneck for the system. The common emitter stage is typically designed with a large amount of degeneration, usually in the form of an emitter resistance. This necessitates an increase in headroom and additional power consumption for the circuit.

Using a balun in place of the active transconductor provides the benefit of perfect linearity without an increase in headroom, and eliminates the power consumption. 

Figure 6.1: Typical differential SEF driven by a transconductor.
of the traditional input stage. If the balun is used to match the switches to the RF input, its differential impedance is fixed; however, the common mode impedance of an ideal balun is considered infinite. This high common mode impedance is desirable when driving the SEF because of the common mode switching action.

If we consider the input source to have a finite resistance $R_S$ we can derive

$$Z_{DM} = \frac{R_S N^2}{2}.$$  

This is derived in Eq. 6.1 assuming that the coupling ‘$k$’ between turns is ideally 1, and that power is conserved. If the balun has a turn ratio

![Figure 6.2: Balun driving point impedance. (a) Differential driving point impedance. (b) Common mode driving point impedance.](image)
of 1, each differential source sees half of the source impedance $R_S$ at their outputs.

$$Z_{DM} = \frac{V_{in}}{I_{in}} \text{ where } I_{in} = \frac{2V_{in}^2}{R_S V_{in} N^2}$$ \hspace{1cm} (6.1)

$$Z_{DM} = \frac{R_S N^2}{2}$$ \hspace{1cm} (6.2)

The common mode impedance at the output of the balun is ideally infinite. This is shown without derivation in Fig. 6.2b where the two currents at the output cancel. As each source produces no output current the illusion of an open circuit is produced, rendering the common mode impedance, $Z_{CM} = \infty$. These are only ideal approximations of the baluns behavior. Any nonidealities will be discussed later as they apply to specific design considerations.

### 6.2 A Single Ended to Differential SEF Utilizing a Balun

This section describes the use of a balun as the passive ‘transconductor’ in a single ended to differential system. Fig. 6.3 represents the system proposed in this section. Using a balun is favorable in a direct sampling receiver design because it produces a differential output, while maintaining the high common-mode impedance which allows the track-and-hold to function. This circuit block could be used in a direct sampling receiver immediately following an LNA. There are downsides to using the single-ended to differential converter which are discussed in the following paragraphs.
A considerable issue in this design is the nonlinear contribution induced by an amplitude or phase mismatch at the input of the track-and-hold. A perfectly balanced single-ended to differential converter is generally impossible to design. Implementing the single-ended to differential converter in the form of a balun results in imbalances across the output terminals of the device. If this imbalanced signal is then fed into the input of the differential track-and-hold and sampled, common mode nonlinearities are not completely rejected. A simple analysis of amplitude imbalance will provide a basis for determining the effect of a non-ideal balun. First a new model for the
balun is shown in Fig. 6.4 where coupling capacitances are modeled to create a more accurate representation of the balun. This highly simplified model depicts the effect of coupling capacitance across the coils of the balun which allow currents to leak from input to output and degrade the ideal performance of the passive.

To quantify the effect of this imbalance, we analyze a pair of emitter followers loaded with hold capacitors using the same theoretical linearity analysis as in Chapter 3. The simple model found in Fig. 6.5 details the operation of the pseudo-differential circuit. The small input signals are defined in Eq. 6.3 and Eq. 6.4 where the amplitude of each signal deviates from a common value by $\pm \delta/2$ as shown in Eq. 6.5 and Eq. 6.6.

\[
\begin{align*}
v_{in+}(t) &= V_{pk+} \cos(\omega t) \\
v_{in-}(t) &= -V_{pk-} \cos(\omega t)
\end{align*}
\]

\[
\begin{align*}
V_{pk+} &= V_{pk} + \frac{\delta}{2} \\
V_{pk-} &= V_{pk} - \frac{\delta}{2}
\end{align*}
\]

The output at the second harmonic frequencies is taken from the Volterra series analysis from Chapter 3. The two output signals at the second harmonic frequency are shown in Eq. 6.7 and Eq. 6.8.
The differential second order distortion $HD_{2-Differential}$ can be quantified by dividing the differential second order signal, by the differential first order signal as shown in Eq. 6.9.

$$HD_{2-Differential} = \left| \frac{\frac{V_{pk+}^2}{2} Re(H_2(j\omega_o, j\omega_o)e^{j2\omega_o t}) - \frac{V_{pk-}^2}{2} Re(H_2(j\omega_o, j\omega_o)e^{j2\omega_o t})}{(V_{pk+} + V_{pk-}) Re(H_1(j\omega_o)e^{j\omega_o t}) + \frac{3(V_{pk+}^3 + V_{pk-}^3)}{4} Re(H_3(j\omega_o, j\omega_o, -j\omega_o)e^{j3\omega_o t})} \right|$$ (6.9)

The second order distortion tends to increase with input imbalance as shown in the equation, and is verified by simulation. The calculated $HD_2$ was compared with a simulated result for the pseudo differential structure and shown in Fig. 6.6. The relative amplitude difference is defined in Eq. 6.10, where $\delta$ is the total difference in amplitude between the two small signals. As the result suggests, if $V_{pk+}$ and $V_{pk-}$ are equal the numerator in Eq. 6.9 vanishes, which is the case for a perfectly balanced differential input. The derived expression can also be used to model the affect of circuit mismatch in the two legs of the differential path and the resulting degradation in $HD_2$. In this case the two amplitudes $V_{pk+}$ and $V_{pk-}$ may match perfectly, but the second order Volterra operators will act differently on the two signal producing an incomplete cancellation of second order distortion.

$$\Delta_{dB} = 20\log_{10} \left( \frac{V_{pk} + \delta/2}{V_{pk} - \delta/2} \right)$$ (6.10)

For very small differences in amplitude, $HD_2$ becomes vanishingly small, as
indicated in the plot. As this amplitude imbalance grows HD$_2$ tends towards a limit dictated by the HD$_2$ of the single ended circuit, under the same operating conditions.

Figure 6.6: A comparison of the simulated and calculated performance of the pseudo-differential circuit, where amplitude mismatch is introduced.

Figure 6.7: A balun simulated in HFSS.

In other words, the performance of the differential circuit can be no worse than its
single ended constituent.

Figure 6.8: S-parameters gathered from simulation of the balun.

With these considerations in mind a balun was designed to perform over a relatively wide band, with a fair degree of amplitude and phase balance. An image of the single-ended to differential converter is shown in Fig. 6.7. The s-parameters for the balun are shown in Fig. 6.8 for two single ended output loads of 25Ωs. The single-ended to differential track-and-hold is depicted in Fig. 6.9. The 100Ω resistor is used to match the transformer to the 50Ω off-chip environment, while maintaining the high common mode impedance necessary to switch the emitter follower. The diode connected device $Q_4$ in each SEF marginalizes the difference in collector voltage on $Q_2$ and $Q_3$ rendering the $V_{CE}$ drop across each device approximately equal. This is also important in ensuring that $Q_3$ does not break down.
6.3 A Differential to Differential SEF Utilizing a Transformer

As it is often difficult to design a single-ended to differential balun which is well matched in amplitude and phase, this section describes the design of a transformer driven differential SEF architecture. By virtue of its differential input, the

Figure 6.9: A single-ended to differential SEF structure.

Figure 6.10: A differential to differential SEF architecture using a transformer as a passive transconductor.
transformer performs better than the balun, both in amplitude and phase mismatch and in bandwidth. The system is depicted in Fig. 6.10 and consists of a transformer, differential SEF, clock driver, two emitter follower buffers, and 50Ω buffer.

Figure 6.11: A 1:1 transformer modeled in HFSS.

The differential SEF is shown in Fig. 6.13. Two feedthrough attenuation capacitors were used to compensate signal leakage through the parasitics of the switch. The 250Ω resistor is used to match the transformer to the 100Ω off-chip environment, while maintaining the high common mode impedance necessary to switch the emitter follower. As mentioned in the previous section, the device \(Q_4\) helps ensure that \(Q_3\) does not break down.
Figure 6.12: S-parameters gathered from simulation of the transformer.

Figure 6.13: A differential to differential transformer based SEF track-and-hold architecture.
The output buffer and clock driver were used in both systems discussed in this chapter. The following two sections briefly discusses the design of these blocks.

### 6.4 Output Buffer

The output buffer was designed to drive the sampled signal off chip, without corrupting the linearity of the signal. The architecture of the buffer is a differential cascoded transconductor, with heavy emitter degeneration. Because of the strict linearity requirements, and cascode structure the supply voltage for the buffer was raised to 6V in order to provide the headroom for the cascode device and degeneration resistor. In total, the buffer draws 39mA from a 6V supply, resulting in roughly 234mW of power consumption. The buffer consumes a majority of the power required by the chip.

![Figure 6.14: Circuit schematic of the 50Ω output buffer.](image-url)
6.5 Clock Driver

The clock driver consists of an input emitter follower, differential amplifier and output emitter follower. The input emitter follower is matched to 50Ω by inserting a resistor from the base of the device to the supply. The differential amplifier is introduced to provide a nominal gain for the clock path. The output emitter follower stage was used to provide low output impedance to the switched emitter follower. In total, the clock driver draws 11.5mA from a 3V supply, leading to a power consumption of roughly 35mW. The bandwidth of the driver is approximately 50GHz with 6dB low frequency gain.

6.6 Performance

This section presents the simulated results of both systems along with images of the chip layouts. The simulated results of the balun driven track-and-hold is
less a performance summary than it is a proof of concept. The results from the transformer driven track-and-hold demonstrates the performance of the system and compares this performance with state of the art high-speed systems. Fig. 6.16 shows the time domain waveform captured at the output of the balun driven system, with its associated spectrum. As the waveform shows, the track-and-hold is capable of switching the output between states, which is evident from the hold pedestal. The spectrum demonstrates the second order distortion due to the amplitude imbalance generated by the balun. These second order tones are located at $2F_1$, $2F_2$ and $F_1 + F_2$.

Figure 6.16: Simulated time domain waveform and output spectrum for the single-ended to differential SEF where $F_{in} = 5.1/5.2\,GHz$ and $F_s = 30\,GS/s$.

A two tone test was used to simulate the linearity of the transformer driven
track-and-hold with input frequencies of 5.1 & 5.2GHz respectively. These frequencies were chosen for this simulation because they are near the lower cutoff frequency of the transformer. The result is shown in Fig. 6.17 where the input power was swept from -25dBm to 12.5dBm, and from the plot its clear that the output begin to saturate. From extrapolation the resulting IIP3 and OIP3 are 22.1dBm and 18.9dBm respectively. This result is compared alongside the performance of other high-speed HBT track-and-holds in SiGe and InP, shown in Table. 6.1. The only circuit in this summary that consumes lower power, is clocked at 18GS/s [37] and if the IIP3 point is extrapolated from their published P1dB result it would be approximately 10dBm, which is more than an order of magnitude lower than this work.

![IIP3 plots simulated for the system, where F_{in} = 5.1/5.2GHz and F_s = 40GS/s.](image)

Figure 6.17: IIP3 plots simulated for the system, where F_{in} = 5.1/5.2GHz and F_s = 40GS/s.
Table 6.1: Track-and-Hold Performance Comparison

<table>
<thead>
<tr>
<th>( F_s )</th>
<th>Linearity ( @ F_{in}, F_s )</th>
<th>Supply(V)</th>
<th>Power</th>
<th>Process</th>
<th>Area(mm(^2))</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>40GS/s</td>
<td>18.9dBm OIP3 @5.1GHz, 40GS/s</td>
<td>+3, +6</td>
<td>330mW</td>
<td>SiGe</td>
<td>1.4 x 0.8</td>
<td>This Work</td>
</tr>
<tr>
<td>40GS/s</td>
<td>22.1dBm IIP3 @5.1GHz, 40GS/s</td>
<td>+3, +6</td>
<td>330mW</td>
<td>SiGe</td>
<td>1.4 x 0.8</td>
<td>This Work</td>
</tr>
<tr>
<td>18GS/s</td>
<td>0dBm P1dB @2GHz, 18GS/s</td>
<td>+3.5</td>
<td>128mW</td>
<td>SiGe</td>
<td>1.6 x 1.7</td>
<td>[37]</td>
</tr>
<tr>
<td>40GS/s</td>
<td>6dBm IIP3 @2.5GHz, 40GS/s</td>
<td>+3.6</td>
<td>540mW</td>
<td>SiGe</td>
<td>1.1 x 1.0</td>
<td>[38]</td>
</tr>
<tr>
<td>40GS/s</td>
<td>15.6dBm IIP3 @10GHz, 40GS/s</td>
<td>+5.5</td>
<td>560mW</td>
<td>SiGe</td>
<td>0.8 x 0.2</td>
<td>[27]</td>
</tr>
<tr>
<td>12GS/s</td>
<td>12dBm IIP3 @12GHz, 12GS/s</td>
<td>-5.2</td>
<td>390mW</td>
<td>InP</td>
<td>.66</td>
<td>[39]</td>
</tr>
<tr>
<td>50GS/s</td>
<td>17.2dBm IIP3 @18GHz, 50GS/s</td>
<td>-5, -2.5</td>
<td>1200mW</td>
<td>InP</td>
<td>.68 x 1.1</td>
<td>[40]</td>
</tr>
</tbody>
</table>

It’s clear that these systems are capable of operating with higher linearity and lower power consumption than other high-speed samplers, but at the cost of narrow band operation and added die area. Both of these disadvantages are inherently introduced by the passive “transconductor”, which limits the variety of applications such a system would be useful for. Although these circuits might not be useful for broadband systems, they still present an advantage in ‘tuned’ applications where multiple gigahertz of bandwidth require sampling. In addition, the use of a passive “transconductor” in an SEF track-and-hold becomes increasingly advantageous at higher frequencies, because the passive device can shrink in area, maybe even approach the area of its active counterpart. For this reason, the track-and-holds introduced in this work are useful for direct sampling of high frequency signals where power consumption is a critical design constraint.
6.6.1 Chip Layouts

This subsection includes layout images of the balun and transformer driven track-and-holds. Fig. 6.18 & Fig. 6.19 show the respective layouts of the balun and transformer driven track-and-holds. Both chips measure 1,400\(\mu\)m by 800\(\mu\)m.

Figure 6.18: Layout of the balun driven track-and-hold chip designed in IBM 9HP.

Figure 6.19: Layout of the transformer driven track-and-hold chip designed in IBM 9HP.
Conclusion

This thesis introduced a number of sampling systems designed in InP and SiGe BiCMOS processes, with varying applications including tests and measurement, optical networking and direct sampling receivers. A thorough discussion of open-loop linearization techniques provided a basis for designing high resolution sampling circuits operating in the 30-40GS/s regime. The circuits presented in this work operate very well when compared with other state of the art systems, achieving higher linearity and lower power consumption in a majority of cases. The architectures described in this work help to advance the state of the art, through innovative circuit techniques which improve upon existing technologies, supporting high bit rate and high frequency applications.
Bibliography


