1

PM-1 3 1/2" x 4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our title Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

Ultralinear & Wideband Integrated Circuit Design

by

John Nisbet, B.A.Sc.

**A thesis submitted to the Faculty of Graduate Studies and Research in
partial fulfilment of the requirements for the degree of**

Master of Engineering

Department of Electronics

**Carleton University
Ottawa, Ontario
January 1993**

**© copyright
1993, John Nisbet**

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-84130-3

Canada

The undersigned recommend to the Faculty of Graduate Studies and
Research acceptance of the thesis

“Ultralinear Wideband Integrated Circuit Design”

submitted by John Nisbet, B.A.Sc.

in partial fulfilment of the requirements for the degree of

Master of Engineering

Department of Electronics

A handwritten signature in black ink, appearing to read 'A. Lyth', is written over a solid horizontal line.

Thesis Supervisor

A handwritten signature in black ink, appearing to read 'J. Nisbet', is written over a solid horizontal line.

Chair, Department of Electronics

**Carleton University
January, 1993**

Abstract

The design of ultralinear wideband integrated circuits is treated from three perspectives: system design concerns, modelling and simulation techniques and circuit design considerations.

The system design phase uses concurrent treatment of noise and distortion to produce an appropriate block-and-level description of hardware partitioning. Conventional SPICE models and analysis can be used for design, but accurate simulations must include modelling refinements such as component mismatches and thermal distortion. The ultralinear design techniques include the use of local feedback and feedforward, current-mode and translinear topologies and a variety of layout techniques which reduce electrical and thermal mismatches.

The design techniques are applied to produce a surface acoustic wave driver amplifier IC with 1/10 the cost, 1/5 the area and 1/4 the power consumption of an equivalent discrete implementation. These techniques can easily be applied to faster bipolar processes to achieve wide bandwidth, superb linearity and low power consumption.

Acknowledgements

This work was only possible because of the support given to me by many people. I would like to thank Al Javed, Bryan Godfrey and the staff of BNR department 1U12, especially Stan Blakey, for their technical and moral support. Many thanks also to my supervisor, Barry Syrett, and to the staff at Tektronix Integrated Circuits Operation for helpful comments and suggestions.

I am very grateful to Karen Westwood and Rick Nowak for their help with document preparation, and to Beyhan Kaya for her help with the computing. The support given by Tom Cameron and Ian McKay to the chip design work is deeply appreciated.

Finally, my heartfelt thanks go to my wife, Heather, for whom the Carleton University motto "Ours The Task Eternal" has acquired unforgettable meaning.

Table of Contents

Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
List of Symbols	xiv
1 Introduction	1
1.1 Scope of Study	1
1.2 Background	4
1.3 Design Tradeoffs for Linear Wideband Circuits	6
1.4 Thesis Organization	7
2 System Considerations.	9
2.1 Introduction	9
2.2 Types of Distortion	9
2.2.1 Linear Distortion.	9
2.2.2 Nonlinear Distortion.	12
2.3 Measures of Distortion.	13
2.3.1 Total Harmonic Distortion (THD).	13
2.3.2 Compression Point.	13
2.3.3 Intermodulation Distortion Products	15
2.3.4 Crossmodulation Distortion	16
2.3.4 Quantization Errors in Digital Circuits	17
2.3.5 Calculation of Distortion Measures	20
2.4 Linearity Issues Specific to QAM Modems	21
2.5 Noise Figure and IMP of Cascaded Amplifiers	33
3 Modelling and Simulation	35
3.1 Introduction	35
3.2 SPICE Models	37
3.2.1 Diffused Resistor Model.	37
3.2.2 MOS Capacitor Model.	39
3.2.3 Bipolar Transistor Model	41

3.3 Simulation Using SPICE	49
3.3.1 Volterra Analysis	50
3.3.2 Differential Gain Analysis.	52
3.3.3 Transient Analysis.	54
3.4 Guidelines for IC Design Using CAD.	58
4 Design Considerations.	60
4.1 Device Design and Operating Point	60
4.2 Single Transistor Stages	66
4.3 Effects of Impedance	69
4.4 Differential Versus Single-Ended Circuits	72
4.5 Effects of Nonideal Bias Circuits	76
4.6 Device Mismatches	86
4.7 Thermal Distortion and Temperature Effects	91
4.8 Feedback.	95
4.9 Feedforward	106
4.10 Translinear and Current-Mode Circuits	109
4.11 Distortion and Noise in Multipliers	117
4.12 Layout Considerations	124
4.13 Applications	128
5 SAW Driver Amplifier	130
5.1 Introduction	130
5.2 Design Requirements	130
5.3 SAW Driver Design	135
5.4 SAW Driver Performance	147
6 Conclusions and Further Work	149
A Distortion Measures	153
Case 1 :.	153
Case 2:	155
Case 3:	157
B Distortion Limits for QAM Systems	162
C QAM Constellation Peaking Factor	166
D Noise and Intermodulation.	169
E The SPICE GP2 Model	174
F Functions of Random Variables.	184
References	186

List of Tables

Table 2.1: Maximum Magnitude and Phase Errors for QAM Signals	25
Table 2.2: Peak-to-rms Ratio for Square QAM with n Bits Per Axis	26
Table 4.1: Output 0.1 dB Compression Power for Single-tansistor Stages	68
Table 4.2: 1 dB Compression Point for Circuits of Figure 4-14.	77
Table 4.3: Distortion versus R_b for Common-Base Amplifier	81
Table 4.4: Distortion for Ideal and Real Current Sinks	86
Table 4.5: Matching Data for SH3 Devices	90
Table 4.6: Distortion as a Function of Emitter Area Mismatch	91
Table 4.7: Thermal Properties of SH3 Devices	94
Table 4.8: Effect of Thermal Modelling on Simulated Distortion.	96
Table 4.9: Dynamic Range Comparison for Transconductances of Figure 4-44	117
Table 4.10: Metal-Substrate Parasitic Capacitance for SH3 Process	129
Table 5.1: PLCC28 Package Properties	140
Table 5.2: Measured SAWD Performance	148
Table 5.3: Effect of Modelling Assumptions on Simulated IM3	149

List of Figures

Figure 2-1 : Linear magnitude and delay distortion	11
Figure 2-2 : Compression point definition.	14
Figure 2-3 : Third-order intercept point definition	16
Figure 2-4 : Multi-level signalling.	18
Figure 2-5 : 8-level quantization of a linear function	19
Figure 2-6 : 8-level and 16-level quantization of a nonlinear function	19
Figure 2-7 : Simple QAM system	21
Figure 2-8 : 16QAM constellation diagram	23
Figure 2-9 : Peak states and constellation boundaries for a square QAM signal.	24
Figure 2-10 : Ideal QAM pulse transmission	27
Figure 2-11 : Pulse responses for zero ISI.	28
Figure 2-12 : Raised cosine frequency response	29
Figure 2-13 : Raised cosine time response	29
Figure 2-14 : Calculation of overshoot caused by overlapping filter pulse responses.	31
Figure 2-15 : Noise figure, gain and n-th order intercept point for a cascade.	33
Figure 3-1 : Cross-section view of a diffused resistor.	37
Figure 3-2 : Model schematic for diffused resistor	38
Figure 3-3 : Diffused resistor carrying dc current showing (a) incursion of depletion layer into resistor body and (b) depletion layer widening	39
Figure 3-4 : Cross-section view of a MOS capacitor	40
Figure 3-5 : Model schematic for MOS capacitor.	40
Figure 3-6 : Cross-section view of a planar NPN transistor	42
Figure 3-7 : Model schematic for planar NPN transistor	42

Figure 3-8 : Distribution of collector-base capacitance across base resistance for the planar NPN transistor	44
Figure 3-9 : Extrapolation of the common-emitter output characteristics to determine the Early voltage	46
Figure 3-10 : NPN model schematic modified to simulate collector-base breakdown	47
Figure 3-11 : Extended Gummel-Poon model derived from basic GP2 model	49
Figure 3-12 : Nonlinear (a) and quasilinear (b) models for Volterra analysis.	51
Figure 3-13 : Replacement of coupling capacitors for differential gain analysis.	54
Figure 3-14 : Frequency and time domains with relations between FFT parameters	56
Figure 3-15 : Incorrect (a) and correct (b) choice of excitation frequency and transform points for FFT.	57
Figure 3-16 : Undesired contribution to measured distortion level by window sidelobe detection of high-level signal	58
Figure 4-1 : Full depletion of collector epi-layer	61
Figure 4-2 : Magnitude and phase of transconductance as functions of collector-base voltage at several values of emitter-base voltage (simulated)	62
Figure 4-3 : Magnitude and phase of transconductance as functions of base- emitter voltage at several values of collector-base voltage (simulated)	63
Figure 4-4 : Transconductance degradation by parasitic base and emitter resistance.	64
Figure 4-5 : Conventional (a) and extended-frequency (b) SPICE GP2 models	65
Figure 4-6 : f_T versus I_C and V_{CE} for an E60X NPN transistor (simulated)	

Figure 4-7 : Single-transistor gain stages used for compression comparison	67
Figure 4-8 : Equivalent Miller capacitances for a voltage amplifier with shunt capacitive feedback	68
Figure 4-9 : Source and load matching for a typical wideband IC	70
Figure 4-10 : Amplifier test circuit	70
Figure 4-11 : Output compression point versus impedance level for the circuit of Figure 4-10	71
Figure 4-12 : Impedance matching of a wideband IC using balun transformers	72
Figure 4-13 : Identical imperfect amplifiers operated in differential mode	73
Figure 4-14 : Single-ended and differential amplifier test circuits	75
Figure 4-15 : Cascode amplifier with non-ideal resistance R_S , conductance G_P and capacitance C_P	77
Figure 4-16 : Emitter impedance of a bipolar transistor.	78
Figure 4-17 : Test circuit to determine the effect of R_S on transresistance linearity	80
Figure 4-18 : Resistor dividers for cascode bias.	81
Figure 4-19 : Collector impedance of a bipolar transistor.	82
Figure 4-20 : Test circuit to determine the effect of real and ideal current sinks on distortion	84
Figure 4-21 : Wilson (a) and cascode (b) current sinks	85
Figure 4-22 : Test circuit to determine the effect of mismatches on $(I_1 - I_2)$	87
Figure 4-23 : Test circuit to determine the effect of mismatches on g_m nonlinearity	90
Figure 4-24 : Thermal model for NPN bipolar transistor	92
Figure 4-25 : Test circuit to determine the effect of thermal distortion on compression and intermodulation	94

Figure 4-26 : Negative feedback topology (a) and its effect on distortion	
(b)	95
Figure 4-27 : Single-transistor amplifiers with series (a) and shunt (b)	
feedback	96
Figure 4-28 : Single-transistor amplifier with combined series-shunt	
feedback	97
Figure 4-29 : Multiple-transistor feedback amplifiers.	98
Figure 4-30 : Single transistor and corresponding three-transistor buffer	
(3TB).	99
Figure 4-31 : Test circuits to compare magnitude and phase compression	101
Figure 4-32 : Magnitude and phase compression for the circuits of Figure	
4-31	102
Figure 4-33 : Voltage feedback amplifier topology	103
Figure 4-34 : Current feedback amplifier topology	104
Figure 4-35 : Feedforward amplifier topology	106
Figure 4-36 : Cascomp amplifier feedforward topology	107
Figure 4-37 : Cascode and cascomp test circuits to be compared for gain	
magnitude and phase compression	108
Figure 4-38 : Magnitude and phase compression for the circuits of Figure 4-	
37	109
Figure 4-39 : Functional implementation of $y = a_1x_1 + Kx_2$	110
Figure 4-40 : Circuit implementation of $y = a_1x_1 + Kx_2$	111
Figure 4-41 : Conventional (a) and folded (b) cascode amplifiers	112
Figure 4-42 : Translinear squaring circuit	113
Figure 4-43 : Hybrid-pi small-signal model of a bipolar transistor including	
noise generators	115
Figure 4-44 : Transconductance noise/distortion test circuits	116
Figure 4-45 : Classical wideband multiplier.	117
Figure 4-46 : Multiplier quad with bias currents at null (a), maximum	
positive gain (b) and maximum negative gain (c).	120

Figure 4-47 : Multiplier test circuit	121
Figure 4-48 : Multiplier current gain for ideal devices (a), emitter area mismatch (b) and parasitic base resistance (c).	122
Figure 4-49 : Multiplier noise test circuit	123
Figure 4-50 : Output noise versus multiplier weight for the circuit of Figure 4-49	123
Figure 4-51 : Axis of symmetry and thermal source line for optimum die layout	124
Figure 4-52 : “Cross-split” technique used to minimize thermal, strain and process offsets	125
Figure 4-53 : Thermal “averaging” used in a ratiometric current multiplier 126	
Figure 4-54 : Emitter metal matching	126
Figure 4-55 : Proper use of dual bondpads and “branching” topology	127
Figure 5-1 : SAW Device	131
Figure 5-2 : SAW device matching strategy.	132
Figure 5-3 : Simplified schematic of discrete SAW driver amplifier	133
Figure 5-4 : Differential cascode SAW driver amplifier topology.	136
Figure 5-5 : SAW driver output current source/sink capability	137
Figure 5-6 : Dynamically-biased SAW driver output current sinks.	138
Figure 5-7 : SAWD amplifier block diagram	140
Figure 5-8 : SAWD amplifier schematic	141
Figure 5-9 : SAWD die floorplan	146
Figure A-1 : Ideal linear gain, real nonlinear gain and gain error for an amplifier	159
Figure A-2 : V_{out} and dV_{out}/dV_{in} for obtaining G_{max} and G_{min}	161
Figure B-1 : Peak state magnitude for various QAM constellations	162
Figure B-2 : Determination of maximum magnitude-only error (e_m).	163
Figure B-3 : Determination of maximum phase-only error (e_p).	164
Figure C-1 : One quadrant of a square QAM constellation	166
Figure D-1 : Cascade of signal processing blocks.	169

Figure D-2 : Input and output signal-to-noise ratio	169
Figure D-3 : Relation between P_{lin} , IP_n and P_{IMn}	171
Figure E-1 : Modified Gummel-Poon model for a bipolar transistor	179

List of Symbols

α	Raised cosine filter rolloff factor OR common-base current gain
β	Common-emitter current gain
γ	Multiplier emitter area mismatch factor
Δf	Frequency bin interval for an FFT
Δt	Time interval in a transient analysis
ϵ_m	Maximum allowable magnitude-only signal error for a QAM signal
ϵ_p	Maximum allowable phase-only signal error for a QAM signal
ϵ_s	Permittivity of silicon
λ_g	Transmission line "guide" wavelength
τ_f	Neutral base transit time (forward or normal active mode)
τ_p	Pole time constant
τ_{th}	Thermal time constant
τ_z	Zero time constant
ψ_j	P-N junction barrier voltage
ψ_{jc}	Collector junction barrier voltage
ω	Natural frequency
A	Signal amplitude
A_E	Emitter area
A_r	Transresistance gain
A_v	Voltage gain
B	Overall two-sided bandwidth of an FFT
BV_{CBO}	Collector-base breakdown voltage, emitter open circuit

BV_{CEO}	Collector-emitter breakdown voltage, base open circuit
c	Transmission line capacitance per unit length
C_C	Collector capacitance
C_μ	Hybrid-pi model collector-base capacitance
C_π	Hybrid-pi model emitter-base capacitance
C_j	P-N junction capacitance
C_{j0}	P-N junction capacitance value at zero bias
CM	Cross modulation
CP()	Compression point
D	Distortion factor
f	Frequency
f_s	Symbol frequency
f_t	Transition frequency (unity short-circuit current gain frequency)
g_π	Hybrid-pi model base-emitter conductance
g_m	Hybrid-pi model transconductance
g_o	Hybrid-pi model output conductance
G_C	Collector conductance
$h(t)$	Time domain impulse response of a linear network
$H(\omega)$	Frequency domain transfer function of a linear network
I_C	Collector current
I_{CH}	Collector current under high-level injection conditions
IM3	Third-order intermodulation level, in dB, below wanted signal level
IP3	Third-order intercept point
I_s	Transport saturation current
J_C	Collector current density

K	Multiplier scaling constant
l	Transmission line inductance per unit length
L_E	Emitter inductance
m	AM modulation index OR P-N junction grading coefficient
M	Avalanche current multiplication factor
n	Number of bits per axis in digital transmission
N	Number of points in an FFT (integral power of two)
N_D	Donor impurity doping concentration
r_b	Hybrid-pi model parasitic base resistance
r_o	Hybrid-pi model output resistance
R_E	Emitter resistance
R_{th}	Thermal resistance
SE	Spectral efficiency of digital modulation
T	Time duration of a transient analysis
THD	Total harmonic distortion
T_s	Symbol period
v_p	Transmission line propagation velocity
V_A	Early voltage
V_{BE}	Base-emitter voltage
V_{CB}	Collector-base voltage
V_{CE}	Collector-emitter voltage
V_D	Depletion voltage
v_t	Thermal voltage; $v_t = kT/q$
W	Nominal information bandwidth of a raised cosine filter
$x(t)$	Time domain representation of variable x

$X(\omega)$	Frequency domain representation of variable x
Y_C	Collector admittance
Z_0	Transmission line characteristic impedance
Z_E	Emitter impedance
Z_{opt}	Transmission line optimum characteristic impedance

1 Introduction

1.1 Scope of Study

The subject of this thesis is distortion in wideband amplifiers and multipliers. Specifically, I will address the system design aspects of distortion, modelling considerations for predicting it and circuit design techniques which reduce it.

Distortion is defined to be a change in the spectral content of a signal, especially when such a change is undesirable. The definition excludes corruption of the signal by additive white gaussian noise (AWGN). Typically, noise sets the lower limit of dynamic range and distortion sets the upper limit. In many applications, such as radar or communication receivers, maximum dynamic range is desired. For this reason, noise performance is as important to the designer as distortion performance.

In principle, wideband circuits should have gain and return loss which are flat with frequency from dc to the edge of the band of interest. In practice, ac coupling is often used, resulting in a low-frequency cutoff greater than zero. The circuits investigated here are qualified as wideband so that we can consider spurious responses across the full pass-band of the circuit. The main application of this study is to monolithic integrated circuits (ICs), which for the most part demand wideband design approaches.

Analog signal processing normally requires a variety of operations to be performed; scaling, summing, multiplication, mixing and so on. This investigation is restricted to those operations where no inherent frequency conversion occurs, especially scaling and summing. Specifically excluded are spectrum spreading and despreading, mixing and frequency multiplication or division. Specifically included are fixed-gain amplifiers,

phase shifters, summing amplifiers and multipliers with a wideband signal input and a DC control input. The multipliers may be either two-quadrant (where gain varies from zero to some maximum value), or four-quadrant (where gain varies from some negative maximum through zero to some positive maximum). The common feature of all these circuits is that a single pure input tone will ideally result in a single pure output tone of the same frequency. This restriction is applied only to simplify interpretation of derived and measured performance; many of the techniques can be applied, for example, to the design of mixers.

This study was undertaken because of a perceived gap between past practice and present capabilities in VHF circuit design. Historically, most design approaches have been oriented toward discrete component design on some type of printed-circuit substrate. Since the early eighties, however, rapid and steady improvement of bipolar silicon IC processes has led to the availability of very high-speed devices, and an increasing awareness of using integrated circuits as analog subsystems to reduce overall cost, size and power consumption. Complex digital modulation schemes, multicarrier transmitters and general crowding of the radio spectrum have created an irresistible demand for higher-fidelity signal processing circuits. The proliferation of portable RF consumer products has provided a strong incentive to design equipment which is smaller, cheaper and more reliable than competing products. Taken as a whole, the situation strongly suggests an intelligent application of analog integrated circuits, with the designer trying to achieve high performance with low power consumption. One goal of the study was to examine design issues for this type of approach; tradeoffs, constraints and guidelines which lead the designer to the best performance compromise.

A reasonable starting point is the block-and-level diagram for the system. When complete, this provides the input and output levels, the noise figure needed and the distortion requirements for each subsystem block. Since there are several ways to specify dis-

tortion performance, the designers must choose the most useful one or rationalize between a variety of specifications. The method for treating noise in a cascaded system is well-known, and a similar type of analysis is useful for treating distortion. I have specifically addressed the linearity concerns for a high-order quadrature amplitude modulation (QAM) system used in a terrestrial digital microwave radio.

With the block-and-level diagram in hand, the designer is prepared to design the actual circuits. In the last ten years, high-frequency integrated circuit design has become completely dominated by computer simulation for three reasons. First, the short design cycles and high costs of product development preclude repeated trial fabrication of designs. Second, the use of transistor array breadboards (the previous mainstay of IC development) is quite risky at VHF frequencies. This is because unmodelled parasitics and distributed parameter effects begin to affect signal transmission from array to array on the breadboard. Third, a simulator can provide node voltages and branch currents with equal ease, without the disturbing effects that real instruments have on physical circuits. Having accepted the importance of simulation, we wish to know what methods of simulation are available, and what models are most suitable.

Finally, with the tools in hand, the process of circuit selection and optimization begins. At this point the issues are more specific; device design, operating points, circuit configurations and so on. It would be hopelessly ambitious to try to collect all the useful circuit design techniques, but it is useful to examine tradeoffs, guidelines and performance bounds.

Using techniques developed in the course of this work, I have designed a monolithic wideband amplifier and compared its performance with simulations using vendor-supplied models. Finally, I discuss some design rules of thumb, some pitfalls and some ideas for further work.

1.2 Background

Interest in amplifier distortion extends back almost sixty years to the watershed papers of Espley [1] and Chaffee [2]. These authors considered a nonlinear transconductance with infinite bandwidth (a triode tube), and a purely resistive load. In each case a variant of the power series method [3] is used to calculate the amplitude of each component of the output spectrum when the input excitation is a single tone. Most standard modern IC design texts [4,5,6] limit their discussion of distortion to this type of analysis. Since these presentations are oriented to hand analysis, normally the exponential transconductance relation is the only nonlinearity considered.

By the mid fifties, bipolar transistors were beginning to appear in low-power, low-frequency applications. By this time it was common practice to analyse the signal chain by modelling with a cascade of linear frequency-dependent networks alternated with nonlinear transconductances of infinite bandwidth. Meyer [7] attempted a first-order correction for frequency by using complex frequency-dependent h parameters. Lotsch [8] and Mialinckrodt and Gardner [9] considered in more detail the nonlinear transconductance (the nonlinear dependence of collector current on base-emitter voltage). Narayanan [10] first used the Volterra series to analyse small frequency-dependent distortions in wideband amplifiers. His model included four nonlinearities (transconductance, avalanche, current-gain and collector-base capacitance). This was a major step forward because the Volterra series method is more accurate and flexible than the power series method. In particular, frequency-dependent nonlinearities and feedback topologies can be considered. However, the analysis is very computationally intensive. Narayanan used the transfer function method, in which the Volterra kernels and their associated transfer functions (usually just the first three) are derived explicitly. The response can then be computed for any desired input within the range of the model. The other Volterra series method, the method of nonlinear currents, was implemented by Nagel [11] in the SPICE circuit analysis program. This method is best suited to larger circuits where identification of the kernels would be

prohibitively difficult. Nagel used the Ebers-Moll model [12], while others [13,14] performed similar analyses using the more accurate Gummel-Poon [15] or charge control model. Variants of the Gummel-Poon model are the most popular in recent simulation.

Besides the frequency-domain methods of power series and Volterra series analyses, the other main methods of simulation are time-domain analysis and harmonic balance analysis. In a time-domain analysis, the nonlinear integro-differential equations representing the entire circuit are solved numerically using a time stepping algorithm, and the results are transformed to the frequency domain using a discrete Fourier transform (DFT). In harmonic balance analysis [16], the circuit is partitioned into two sections; one containing all the linear circuit elements and one containing all the nonlinear elements. The linear circuit is solved in the frequency domain, the nonlinear circuit in the time domain, and an iterative DFT algorithm attempts to match up these solutions at the interface between the two parts of the circuit. Frequency-domain methods are most useful for weak nonlinearities [17], while time-domain methods are slower but handle strong nonlinearities more accurately. The hybrid method, harmonic balance, is an attempt to get the accuracy of time-domain solutions with the speed advantage of a frequency-domain simulation.

Relatively few authors have considered the effect of transistor design on distortion. D'Altroy et. al [18] described the design of discrete NPN power transistors for use in a wideband coaxial-carrier system. A much more rigorous treatment by Abraham and Meyer [19] provided design guidelines for planar epitaxial devices, as did Poon [20]. This information is directly useful for either optimizing a high-speed bipolar process, or for calculating the best performance attainable with a given process. Most vendors optimize their performance for high transition frequency (the frequency at which current gain is projected to drop to unity), and for low noise if the process is specifically intended for analog circuits. These parameters are most strongly affected by base-layer and emitter-base junction properties, while linearity is also strongly affected by collector-base junction

properties and bias conditions. Normally the designer must choose the size of each device in his or her circuit based on a consideration of noise figure and distortion performance together.

It is neither possible nor useful to list comprehensively all the various circuit approaches to wideband linearity which have appeared in the literature. Some of the more notable circuit approaches to the wideband amplifier problem are from Meyer et al. [21], Chan and Meyer [22], Thomas [23] and Gilbert [24]. The wideband multiplier problem was addressed by Gilbert [25] and Sansen and Meyer [26]. Since most of my research is aimed at finding circuit solutions to the wideband linearity problem, this subject will be addressed in more detail later.

1.3 Design Tradeoffs for Linear Wideband Circuits

I will conclude this introduction with a brief look at some of the tradeoffs required for the design of linear wideband circuits. The most basic tradeoff for any wideband circuit design is between bandwidth and power consumption. To achieve the highest possible transition frequency, bipolar transistors must be operated with large transconductances and low parasitic capacitive admittance. This implies high emitter current densities, and consequently a relatively large current drawn from the power supply rails. Choice of supply voltages depends on the process breakdown voltages and on the "headroom" needed for the circuit topology chosen. There is also a tradeoff between power density and reliability.

Demanding good linearity from a wideband circuit usually complicates these tradeoffs. Power consumption is increased even further because each device must be biased with high emitter current and collector-base voltage. Higher currents also mean a higher shot noise contribution to overall noise figure, transconductance degradation and aggravated thermal reliability problems. Under large-signal conditions, the instantaneous operating point of a transistor deviates significantly from its quiescent point. This change

causes a time-varying power dissipation in the device, and time-dependent thermal gradients are set up. These temperature variations cause device parameters to vary, and the result is thermal distortion. Thermal distortion can be the limiting factor for ultralinear circuits.

Integrated circuits frequently make use of differential topologies, translinear circuits and other techniques which rely on good device-to-device matching. The effects of device mismatches can also dominate distortion performance, and must be accounted for statistically.

1.4 Thesis Organization

Most of the work was oriented towards circuit design, and the organization of the thesis reflects that.

Chapter 2 details the general system design issues that any designer must consider in order to provide the most cost-effective solution to the overall requirements. Special attention is paid to the design requirements for high-order QAM systems because this is the primary area of application for ultralinear ICs.

In Chapter 3, the issues of modelling and simulation are treated. The early phase of the work was devoted to finding the most effective simulation technique for a good balance of speed and precision. Modelling refinements which are necessary for simulating ultralinear circuits are also discussed.

Most of the design work is discussed in Chapter 4. This is an examination of possible sources of nonlinearity in wideband ICs, and an assessment of their simulated and actual effects on circuit performance. There is also some discussion of the best circuit design and layout techniques to use for these types of circuits.

Chapter 5 deals with a practical application of these techniques. A SAW driver am-

plifier IC was specified, designed and fabricated. The design details are discussed, and performance measurements are presented. The effects of various design assumptions are related to the material in Chapter 4.

Finally, Chapter 6 summarizes the conclusions of the work and suggests new avenues of investigation and exploitation.

2 System Considerations

2.1 Introduction

In developing a communication or radar system, a properly executed system design phase provides the greatest opportunity to achieve high performance at low cost. To the circuit designer, the most useful product of a successful system design is a "block and level" diagram of the analog signal processing chain, where the blocks represent real physical circuits to be designed. In a complete block and level diagram, the nominal signal level (and range of signal levels if applicable) is defined along with the impedance at the input and output of each block, and each block is defined in terms of its passband frequency limits, gain, noise figure and some linearity measure. Normally the system designers make a first pass at the block diagram, then the system designers and circuit designers perform an iterative modification of the design together. The final design usually achieves a compromise between cost and performance. The remainder of this chapter is concerned with aspects of system design that affect the circuit designer, particularly with regard to distortion. The various types of distortion are first defined and the distortion measures which are useful in communication systems are discussed. After touching upon the power series treatment of nonlinear circuits, the linearity requirements for quadrature amplitude modulation (QAM) systems will be presented.

2.2 Types of Distortion

There are two broad types of distortion; linear and nonlinear.

2.2.1 Linear Distortion

Linear distortion or dispersion is the simplest type since it can be represented by a

single transfer function [27]. If we consider a circuit whose input signal (voltage or current) is $x(t)$ in the time domain or $X(\omega)$ in the frequency domain, and whose corresponding output signal is $y(t)$ or $Y(\omega)$, then distortionless transmission requires that the output be a scaled and delayed replica of the input:

$$y(t) = Ax(t - t_0) \quad (2.1)$$

Taking the Fourier transform of Equation (2.1) yields:

$$Y(\omega) = AX(\omega) \exp(-j\omega t_0) \quad (2.2)$$

Defining the transfer function in the frequency domain to be $H(\omega)$ we obtain:

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} = A \exp(-j\omega t_0) \quad (2.3)$$

Thus, the requirements for distortionless transmission imply a transfer function whose magnitude is constant over all frequencies of interest, and whose phase is proportional to frequency over all frequencies of interest. This is often referred to in practice as a "flat" or "linear phase" transfer function. The constant A is the gain or loss of the circuit, and the constant t_0 is the group delay or phase slope. Since, according to Equation (2.3), distortionless transmission requires flat magnitude and linear phase characteristics, it follows that any transfer function that does not satisfy both of these conditions does introduce distortion. This distortion is the linear distortion mentioned above, and its presence may be accounted for by allowing the magnitude and phase of the transfer function to be more general functions of frequency:

$$H(\omega) = A(\omega) \exp(j\Phi(\omega)) \quad (2.4)$$

Since the distortion is linear, there is a single transfer function and the full range of circuit analysis techniques may be used, including Fourier transforms and superposition.

In practice, linear distortion can be compensated for by using equalizers. Since gain magnitude and group delay will ideally be constant with frequency over the band of interest, the linear distortion specifications most often used are the peak-to-peak variations in gain magnitude, ΔM , or group delay, ΔD , as shown in Figure 2-1.

Figure 2-1 : Linear magnitude and delay distortion

2.2.2 Nonlinear Distortion

Linear distortion is characterized by a transfer function which varies with frequency. The other type, nonlinear distortion, is characterized by a transfer function which varies with amplitude. This type of distortion is typical of active devices such as diodes and transistors, and its presence greatly complicates circuit analysis. In general, the consequences of nonlinear distortion are (1), the output contains frequencies which are not present at the input and (2), the output power at a given frequency is not proportional to the input power at that frequency.

The simplest special case of nonlinear distortion is memoryless distortion, where the output signal is an instantaneous function of the input signal. For this type of distortion, the transfer function depends on amplitude only, and not on frequency. Physically, this corresponds to a circuit which contains only linear and nonlinear resistances or conductances, as well as controlled sources which may be linear or nonlinear transresistances or transconductances. An amplifier operating well below its first pole frequency is often treated as having memoryless distortion. In practice this means that any reactive elements present can be neglected in comparison to the resistive elements. Since the circuit variables (node voltages and branch currents) depend only on the instantaneous values of other circuit variables, any circuit function can be represented using a power series.

Besides the special cases of linear distortion (dependent on frequency only) and memoryless distortion (dependent on amplitude only), we have the more general case of nonlinear distortion which depends on both amplitude and frequency. Although the analysis for this type of distortion is much more demanding than for the other two cases, it is really the most interesting case because it models by far the largest body of circuits, including wideband integrated circuits. This is because typical semiconductor devices are modelled with nonlinear capacitances as well as nonlinear conductances and transconductances.

2.3 Measures of Distortion

Discussion of the details of modelling will be delayed until Chapter 3 because these details are not usually considered in the system design phase. The crucial point in system block and level definition is the need for a mutually agreed definition and specification for distortion used by both the system designers and the circuit designers. Unlike parameters such as gain and noise figure, many different specifications exist for distortion. Five of the most useful specifications are discussed, then the measures are rationalized in Appendix A.

2.3.1 Total Harmonic Distortion (THD)

Consider a slightly nonlinear amplifier with a single tone input $v_x(t) = V_x \cos(\omega_1 t)$. The output spectrum is a set of harmonics of ω_1 ; the desired output is $V_{y1} \cos(\omega_1 t + \phi_1)$, the second harmonic component is $V_{y2} \cos(2\omega_1 t + \phi_2)$ and so on for higher harmonics. One useful specification for distortion in a wideband amplifier is total harmonic distortion. This is the ratio of the power sum of distortion components at $2\omega_1$, $3\omega_1$ and so on to the power of the desired output [28]:

$$THD = \frac{V_{y2}^2 + V_{y3}^2 + \dots}{V_{y1}^2} \quad (2.5)$$

Distortion factor D is the square root of the total harmonic distortion. THD can be expressed per unit as in Equation (2.5), in percent or in dB. This distortion measure generally varies with input frequency and level.

2.3.2 Compression Point

The compression point is a more useful measure of amplifier linearity because it includes the input level variation. Suppose a graph of output power at ω_1 versus input power at ω_1 is drawn as in Figure 2-2, (both axes have units dBm).

Figure 2-2 : Compression point definition

An ideal amplifier will be characterized by a straight line with unit slope whose y-intercept is the gain of the amplifier in dB. Real amplifiers, however, are characterized by a plot which droops downward at high input power levels. This behavior is known as saturation or compression, and it leads to a convenient and popular specification of the power-handling capability of the amplifier; the 1-dB compression point. Referring again to Figure 2-2, the 1-dB compression point is the signal level at which the actual output power at ω_1 is below the linearly-extrapolated power by 1 dB [29]. The "ideal" line for the amplifier is found by measuring or simulating gain at a very low power level to obtain one point in the (P_{in}, P_{out}) plane, then constructing a line with unit slope through this point. Compression point power may be referred either to the input or output of the circuit. Output power is the most commonly specified.

Since we are only interested in one frequency at the input and output ports, we can measure the output coherently, making possible a measurement of both amplitude and

phase. This is the principle behind the "power sweep" option available on modern network analyzers: a measurement of gain magnitude or phase as a function of input level at some constant frequency. For sophisticated digital modulation schemes such as high-order QAM, phase compression can be as detrimental as amplitude compression. Sometimes other compression points are used, such as 0.1 dB or 1 degree. The IC design community more usually uses percent compression rather than dB.

2.3.3 Intermodulation Distortion Products

If a single tone at frequency ω_1 excites a nonlinear circuit, the output will generally contain components at $m\omega_1$, where $m = \{0, 1, 2, \dots\}$. If there are two excitation tones at ω_1 and ω_2 , the output may contain tones at $m\omega_1 + n\omega_2$, where $m, n = \{\dots, -2, -1, 0, 1, 2, \dots\}$. This spectrum will obviously include harmonics of ω_1 ($n=0$) and ω_2 ($m=0$), as well as other products where neither m nor n is zero. These other components are known as intermodulation products, and their presence is frequently of great interest to the designer.

Suppose an amplifier is excited by two tones closely spaced in frequency, say 99 MHz and 100 MHz. If the amplifier is nonlinear, the output can contain components with $(m=2, n=-1)$ and $(m=-1, n=2)$. These products would be at 98 MHz and 101 MHz. These two intermodulation products are closest in frequency to the desired output frequencies, and are called third-order intermodulation products because the order $(|m| + |n|)$ is three. Since the products fall into the band of interest, no relief can be had by extra filtering, and we must ensure that the level of these third-order products is sufficiently low for all desired output levels and frequencies.

The graph of output power versus input power is useful for characterizing third-order products as well as the fundamental. We first plot the ideal and actual transfer curves for the fundamental, then the ideal and actual curves for the third-order power as shown in Figure 2-3.

Figure 2-3 : Third-order intercept point definition

If we ignore for the moment the fact that both the fundamental and third-order outputs experience compression, we note that the ideal linear extrapolations intersect in a point called the third-order intercept point. This is the point $(P_{in}(IP3), P_{out}(IP3))$ in Figure 2-3 [30]. Other intercept points can be defined for second-, fourth-, fifth- and higher-order products. For bandpass applications, the third-order intercept is the most important, while for lowpass (wideband) applications both second-order and third-order intercepts are of interest.

2.3.4 Crossmodulation Distortion

Another distortion measure frequently used to specify multicarrier systems is cross-modulation. If an AM-modulated carrier at ω_1 and an unmodulated carrier at ω_2 are input to a nonlinear amplifier, the output component at ω_2 will have some unwanted modulation applied to it [31]. For memoryless circuits the cross-modulation factor CM is de-

defined as the ratio of transferred fractional modulation present on ω_2 at the output to the original fractional modulation present on ω_1 at the input; it is a power ratio. CM for these circuits is directly related to the third-order intermodulation IM3 (the power ratio of third-order distortion products to fundamental signals at the output). In fact, as shown in Appendix A:

$$\frac{CM}{IM3} = 4 \quad (2.6)$$

For nonlinear circuits with memory, the situation is more complicated because amplitude modulation at the input (on the carrier at ω_1) is converted to both amplitude and phase modulation on the output carrier at ω_2 . There is no such simple relationship as Equation (2.6) for this case, although some insight can be gained by using more sophisticated analytical techniques as outlined in Chapter 3.

2.3.4 Quantization Errors in Digital Circuits

One other measure of linearity is especially relevant to multi-level digital communication systems, and to QAM systems in particular: quantization linearity. Multi-level signalling schemes inherently rely on the synchronous conversion of parallel bit streams to a single stream of symbols at the transmitter, and a reverse conversion at the receiver as shown in Figure 2-4.

Figure 2-4 : Multi-level signalling

The rate f_s at which the digital-to-analog and analog-to-digital conversions are performed is the symbol rate or baud rate, and the bit rate is nf_s . We can imagine the ideal quantization of the continuous ADC input to a discontinuous output as shown in Figure 2-5. For the perfectly linear transfer curve shown, the familiar equal-step staircase function minimizes mean square quantization error. If the equal step staircase is used to quantize a nonlinear function, as shown in Figure 2-6, the "outer" states (values at the extremes of the input range) may be incorrectly quantized by a staircase which is correct for the "inner" states (values near the center of the input range), if enough bits of resolution are chosen.

For the example shown, the input would be correct if quantized to three bits (eight levels), but incorrect if quantized to four bits (sixteen levels) because in the latter case the outer states at each end of the range are never reached. Thus there is an inherent "floor" of resolution in bits, defined as the maximum number of bits for which all values of the input are correctly quantized. Since the channel and analog signal processing circuits are frequency dependent, a useful specification is the number of bits of quantization at a given quantization rate and for a given input bandwidth. This specification is obviously strongly related to the gain magnitude and phase compression specification, and is crucial to the designer of the wideband amplifier which typically drives the ADC.

Figure 2-5 : 8-level quantization of a linear function **Figure 2-6 : 8-level and 16-level quantization of a nonlinear function**

2.3.5 Calculation of Distortion Measures

As shown in Appendix A, a knowledge of the circuit transfer function coefficients allows any of the distortion measures discussed above to be calculated. For the memory-less amplifier whose output $y(t)$ is given in terms of its input $x(t)$ as:

$$y(t) = a_1 x(t) + a_2 x^2(t) + a_3 x^3(t) \quad (2.7)$$

the total harmonic distortion with single tone excitation $x(t) = A \cos(\omega t)$ is:

$$THD = \frac{\left(\frac{a_2 A^2}{2}\right)^2 + \left(\frac{a_3 A^3}{2}\right)^2}{\left(a_1 A + \frac{3}{4} a_3 A^3\right)^2} \quad (2.8)$$

The input amplitude A for 1 dB output compression is:

$$A = \sqrt{0.1450 \left| \frac{a_1}{a_3} \right|} \quad (2.9)$$

For the two-tone input $x(t) = A(\cos(\omega_1 t) + \cos(\omega_2 t))$, the input amplitudes for second-order and third-order intercept points are:

$$A_{IP2} = \left| \frac{a_1}{a_2} \right| \quad (2.10)$$

$$A_{IP3} = \sqrt{\frac{4}{3} \left| \frac{a_1}{a_3} \right|} \quad (2.11)$$

For the (AM+CW) excitation $x(t) = A(1 + m \cos(\omega_m t)) \cos(\omega_1 t) + \cos(\omega_2 t)$, the cross-modulated output at $\omega_2 \pm \omega_m$ has an amplitude:

$$CM = 3a_3mA^3 \quad (2.12)$$

For a signal constrained to the input range $(-X_{\max}, +X_{\max})$, the worst-case quantization linearity in bits is:

$$n = \text{int} \left(1 + \log \left(\frac{a_1}{a_2 X_{\max} + a_3 X_{\max}^2} \right) \right) \quad (2.13)$$

where the logarithm has base two.

2.4 Linearity Issues Specific to QAM Modems

The specific interest at the time of this work was the development of wideband integrated circuits for IF signal processing in high-order QAM modems. With this in mind, some linearity issues that apply specifically to QAM systems will now be examined.

Figure 2-7 : Simple QAM system

Figure 2-7 shows the block diagram for the most basic type of QAM system. In the modulator shown, $2n$ bits of parallel digital data are converted to two analog baseband channels by two n -bit DACs at a symbol rate f_s . These channels are low-pass filtered to remove spectral foldover and spurious tones, then modulated in quadrature up to an IF carrier frequency. The two IF signals are summed and bandpass filtered, then transmitted. In the demodulator, the IF signal is split and demodulated in quadrature to bring it back to baseband. The mixer outputs are low-pass filtered to remove spurious tones and aliases, and each channel is converted to n parallel digital bit streams by an n -bit ADC. With this system, there are $2n$ bits per symbol and f_s symbols per second. The IF bandwidth occupied by the signal is ideally f_s (in practice less than f_s due to bandpass filter constraints). The spectral efficiency SE of such a signalling scheme is defined as the number of bits per second in the channel divided by the channel bandwidth:

$$SE_{ideal} = \frac{2nf_s}{f_s} = 2n \quad (2.14)$$

One convenient method to visualize the operation of a digital modulation scheme such as QAM is to use a constellation diagram. This is a rectangular plane depiction of the signal space, with one axis representing the I (inphase) signal and the other representing the Q (quadrature) signal. Figure 2-8 shows the constellation for 16-QAM modulation (two bits for each of the I and Q axes). Each dot or "star" in the constellation represents one symbol; for n bits per axis there are 4^n dots. The dots are uniformly spaced in both directions, as are the decision boundaries which fall midway between the dots.

Figure 2-8 : 16QAM constellation diagram

If the signal constellation is considered to be a vector field, then the linear operations of scaling and summing constitute a scalar multiplication on this field, and the resultant constellation should always be a scaled replica of the original constellation. Under this condition, angles are preserved and distances are proportional to the scaling constant. We can use this information to calculate the required magnitude and phase linearity to transmit QAM faithfully.

Let the limits of the signal space be ± 1 on each axis as shown in Figure 2-9, and let the signal space be quantized with n bits per axis (2^n states in each direction). Each dot or state lies at the center of a square decision space, and for a simple square constellation the peak states are the four states at the extreme diagonal corners of the square.

Figure 2-9 : Peak states and constellation boundaries for a square QAM signal

The ideally scaled signal vector would start at the origin of the constellation and terminate on the point in the center of the decision space. We can consider the consequences of magnitude-only (ϵ_m) and phase-only (ϵ_p) errors which cause the signal to deviate from the center to the limits of the decision boundary (Appendix B):

$$\epsilon_m = \frac{100}{2^n - 1} \quad (2.15)$$

$$\epsilon_p = \text{atan} \left(\frac{\left(\sqrt{1 - 2^{1-2n}} \right)}{1 - 2^{1-n}} \right) - 45 \quad (2.16)$$

Equations (2.15) and (2.16) specify the required intrinsic linearity of a circuit handling a QAM signal. Table 2.1 below shows sample values of intrinsic magnitude and phase linearity for square QAM constellations of n bits per axis.

Table 2.1: Maximum Magnitude and Phase Errors for QAM Signals

n	points	ϵ_m (%)	ϵ_p (deg)
1	4	-	45.00
2	16	33.33	16.87
3	64	14.29	7.69
4	256	6.67	3.70
5	1024	3.23	1.82

It is important to remember that these are error boundaries; that is, if the errors are equal to or greater than those specified above, faithful transmission is impossible. The design values of magnitude and phase linearity must be considerably smaller than ϵ_m and ϵ_p .

The actual performance demanded is higher than expected for two reasons: constellation shape and filter peaking. Both these factors increase the peak-to-rms (root mean square) levels in the circuit, and place more stringent demands on analog signal processing elements. Implicit in the definitions of gain magnitude and phase compression and two-tone intermodulation is the idea of sinusoidal input and output signals. The peak-to-rms ratio for a sinusoid is 1.4142 or 3 dB, but this is not the case for a QAM signal. The issue of peak-to-rms ratio is therefore important, since system design and measurements are performed in an rms power context, but the physical elements in the system must handle the peak power with adequate linearity.

The constellation peak-to-rms relationship is simple but tedious to derive (Appendix C). For a classical square constellation with n bits per axis we have:

$$\left(\frac{pk}{rms} \right)_{const} = \sqrt{\frac{2^{3n-1} - 2^{2n} + 2^{n-1}}{\sum_{i=1}^{2^{n-1}} (2i-1)^2}} \quad (2.17)$$

Table 2.2 below provides sample values of constellation peak-to-rms ratio for square QAM with n bits per axis.

Table 2.2: Peak-to-rms Ratio for Square QAM with n Bits Per Axis

n	points	(pk/rms)	dB
1	4	1.0000	0.00
2	16	1.3416	2.55
3	64	1.5275	3.68
4	256	1.6270	4.23
5	1024	1.6787	4.50

In effect, the level handled by the analog signal processing circuits is increased by the amount shown in the table. Sometimes digital mapping is used to transform the standard square constellation into another shape with a lower peak-to-rms ratio. In this case, Figure 2-7 is altered by the addition of a $2^n \times n$ bit ROM (read-only memory) before each DAC in the modulator, and a complementary $2^n \times n$ bit ROM after each ADC in the demodulator.

Filtering also has a pronounced effect on the peak-to-rms ratio of the QAM signal. In any practical communication system, filtering is needed to strictly band-limit the information-carrying signal so that it will not interfere with communications in adjacent channels, and to define the noise bandwidth. In a digital communication system where the information is carried by a train of pulses, the filtering will shape the pulses and the filter characteristics must be designed so that the responses to individual pulses do not overlap at the filter output. In other words, the filter must not cause intersymbol interference (ISI) [32].

Figure 2-10 : Ideal QAM pulse transmission

Figure 2-10 shows ideal QAM pulse transmission; the input $x(t)$ is a series of non-return-to-zero (NRZ) square pulses evenly spaced with period T_s ($f_s = 1/T_s$ = symbol rate). The filtered output $y(t)$ will be sampled at the rate f_s , and the value of the output is only of interest at the sampling instants as far as data recovery is concerned. Thus, one obvious way to remove ISI is by employing a filter time response which has a peak at the sampling instant and which is zero-valued outside $\pm T_s/2$ from the sampling instant: a strictly time-limited signal as shown in Figure 2-11a. Unfortunately, it is not possible to design a realizable filter whose response is both strictly time-limited and strictly band-limited [33].

Figure 2-11 : Pulse responses for zero ISI**(a) unrealizable pulse response** **(b) realizable pulse response**

A useful class of filters exists with strict band-limiting and time responses which have a peak at one sampling instant and are zero-valued at all other sampling instants as shown in Figure 2-11b. These are Nyquist filters [34], and their most common realization is the raised cosine characteristic [35]. The frequency-domain characteristic is of the form:

$$= T_s, |\omega| \leq (1-\alpha)W$$

$$Y(\omega) = \frac{T_s}{2} \left(1 - \sin \left((|\omega| - W) \frac{\pi}{2\alpha W} \right) \right), (1-\alpha)W \leq |\omega| \leq (1+\alpha)W \quad (2.18)$$

$$= 0, |\omega| \geq (1+\alpha)W$$

This is plotted in Figure 2-12. The parameter α is called the roll-off factor; the smaller α is, the steeper the roll-off of the filter skirt ($0 < \alpha < 1$). The nominal information bandwidth is W , while the channel bandwidth is $(1 + \alpha)W$, so a small value of α leads to efficient

use of the available channel bandwidth.

Figure 2-12 : Raised cosine frequency response

The time-domain transform of Equation (2.18) is [34]:

$$y(t) = \left(\frac{\sin(Wt)}{Wt} \right) \left(\frac{\cos(\alpha Wt)}{1 - \left(\frac{2\alpha Wt}{\pi} \right)^2} \right) \quad (2.19)$$

As shown in Figure 2-13, this function has zero value at $t = nT_s$, where n is an integer not equal to zero, so ISI is ideally eliminated at the sampling instants.

Figure 2-13 : Raised cosine time response

This is the minimum requirement for functional communication, but the analog signal processing circuits must also handle the signal with good linearity between sampling instants. It turns out that the signal amplitude between sampling instants can be appreciably higher than the magnitude of the peak state, a phenomenon described as overshoot. It is fairly easy to make an estimate of this overshoot for a given filter.

Consider the case where the filter output is sampled at $t=0$ while the system is receiving the maximum positive signal, shown normalized to 1.0 in Figure 2-14. This response is $y_0(t)$. Suppose the following symbol, one sampling instant later, is also at the peak positive state; this is $y_1(t)$ in Figure 2-14. Then suppose the preceding symbol $y_{-1}(t)$ was at the peak negative state in Figure 2-14. That is:

$$y_0(t) = y(t) \quad (2.20)$$

$$y_1(t) = y\left(t - \frac{\pi}{W}\right) \quad (2.21)$$

$$y_{-1}(t) = -y\left(t + \frac{\pi}{W}\right) \quad (2.22)$$

Figure 2-14 : Calculation of overshoot caused by overlapping filter pulse responses

Now consider the magnitude of the signal at $t=\pi/(2W)$, midway between the sampling instants 0 and 1, for this combination of three symbols passing through the system:

$$y(t) = y_{-1}(t) + y_0(t) + y_1(t) = \frac{2}{3\pi} \left(\frac{\cos\left(-3\pi\frac{\alpha}{2}\right)}{1-9\alpha^2} \right) + \quad (2.23)$$

$$\frac{2}{\pi} \left(\frac{\cos\left(\pi\frac{\alpha}{2}\right)}{1-\alpha^2} \right) + \frac{2}{\pi} \left(\frac{\cos\left(\pi\frac{\alpha}{2}\right)}{1-\alpha^2} \right)$$

The maximum overshoot value is for $\alpha=0$, which is a physically unrealizable "brick wall" filter but which provides a useful bound on the result. If Equation (2.23) is evaluated with $\alpha=0$, the result is 1.4854 corresponding to an overshoot of 3.44 dB. If more terms are added corresponding to the "worst case" polarities of interfering symbols, the peaking be-

comes more pronounced. If we consider the worst case α to be zero and extend Equation (2.23), we have:

$$y(t) = \dots + y_{-1}(t) + y_0(t) + y_1(t) + \dots = \frac{4}{\pi} \sum_{n=0}^{\infty} \frac{1}{2n+1} \quad (2.24)$$

The series in Equation (2.24) does not converge to a limit, so in theory the peak signal amplitude is unlimited! In fact, for $\alpha > 0$ the series does converge, and in practice data scrambling and error control coding can help reduce the probability and severity of transmission errors due to nonlinearity. The need for error control coding is described below, but first consider the case where five symbols on either side of $t = \pi/(2W)$ arrive in their worst case signs and magnitudes. The signal magnitude will be:

$$y_{\max} = (4/\pi)(1 + 1/3 + 1/5 + 1/7 + 1/9) = 2.2757 \quad (2.25)$$

In this case, the ratio of the peak overshoot amplitude to the peak state amplitude is 7.14 dB.

Error control coding ensures that transmission errors caused by distortion due to excessive overshoot can be corrected if they are sufficiently rare. In the case above where we considered a total of ten symbols (five on each side of $t = \pi/(2W)$), and we again consider n bits per axis, the worst case combination of symbols and its negative complement are two cases out of $(2^n)^{10}$ total combinations. If $n=3$ (a 64-QAM square constellation), the total number of combinations is 2^{30} or more than one billion (10^9). If all combinations are equally probable and the two worst case states consistently cause errors, the background uncorrected error rate is $2/2^{30}$ or 1.8626×10^{-9} .

To summarize, the power handling requirements for circuits in QAM systems are defined by the rms power plus two peaking factors; one for the constellation shape and the other for the filter overshoot. In calculating signal magnitudes, the rms power can be

converted to rms voltage or current, then to peak voltage using either 3 dB peak-to-rms (for sinusoidal test tones) or the sum of filter and constellation peaking (for the QAM signal). For example, a 256QAM square constellation with $\alpha=0$ and ± 5 symbol worst case overshoot, the total peak-to-rms is 4.23 dB (constellation) plus 7.14 dB (filter) for a total of 11.37 dB. This is a worst-case bound, but it does demonstrate that linearity requirements for QAM systems are considerably more stringent than for constant envelope modulation.

2.5 Noise Figure and IMP of Cascaded Amplifiers

One final system-related topic concerns the behavior of cascaded blocks with respect to noise and intermodulation. The overall noise figure of a cascade in terms of the noise figures and gains of its component blocks is widely understood and used in communication system design.

Figure 2-15 : Noise figure, gain and n-th order intercept point for a cascade

For the cascade in Figure 2-15, where the i th block has noise figure F and power gain G , the overall noise figure is (Appendix D):

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \dots + \frac{F_k - 1}{G_1 G_2 \dots G_{k-1}} \quad (2.26)$$

A similar relation can be written for the n th-order intercept point of the cascade in terms of the n th-order intercepts of the component blocks (Appendix D):

$$IP_n = \left(\frac{IP_{n_1}^{\frac{1-n}{2}}}{(G_2 G_3 \dots G_k)^{\frac{n-1}{2}}} + \frac{IP_{n_2}^{\frac{1-n}{2}}}{(G_3 G_4 \dots G_k)^{\frac{n-1}{2}}} + \dots + IP_{n_k}^{\frac{1-n}{2}} \right)^{\frac{2}{1-n}} \quad (2.27)$$

The noise figure contribution of any stage to the overall noise figure is divided by the gain preceding that stage, but the intermodulation contribution of any stage to the overall intermodulation is divided by a function of the gain following that stage. For this reason, the gain profile of the signal chain must be intelligently chosen to minimize noise figure and maximize intercept points with an implementation of reasonable cost and power consumption.

Equations (2.26) and (2.27), together with the power budget, are the essential core of a signal chain optimization. The mechanics of this optimization are beyond the scope of this thesis, although an intuitive cut-and-try approach using a computer spreadsheet is one very simple method that works well in practice. The crucial issue at this point is how to accurately model the noise and distortion performance of the blocks at the circuit level.

The choice of distortion measure for any given circuit block is usually determined by its position in the receiver chain. As a rule of thumb, any signal processing blocks which precede the final channel-selection filter will be specified using intermodulation or cross-modulation measures because the adjacent channel signals are still present with the wanted signal there. Once the final channel filtering has been performed, only the wanted signal is theoretically present, so the blocks following the filter are specified in terms of amplitude and phase compression.

3 Modelling and Simulation

3.1 Introduction

As mentioned in Chapter 2, it is best to optimize the performance/cost ratio in the system design phase of the design cycle. In order to realize the expected subsystem performance, careful attention must be given to the modelling and simulation phase.

Modelling and simulation are conceptually separate activities. Modelling is the process of describing the physical circuit with a simplified abstraction, usually using an electrical equivalent circuit. Simulation is the process of "solving" the described circuit to yield the circuit variables (node voltages and branch currents). Since modelling is a prerequisite for simulation it will be discussed first.

Three of the issues relevant to modelling ultralinear circuits are those considered for any type of modelling: accuracy, complexity and basis. Accuracy is a measure of how well a model predicts the observed physical behaviour of a circuit. For reliable simulation, circuit designers seek models which properly predict a wide range of behavioural phenomena over a wide range of operating conditions. The accuracy of a model can be quantified in terms of error between predicted and measured performance over a specified range of operating conditions. Normally several models are used for each device in the circuit in the course of design; for example, different models may be used for bias analysis, noise analysis and distortion analysis. This is done to provide accuracy without using unduly cumbersome models.

Complexity is affected by two criteria: the number of elements used in the model to describe the physical circuit, and the sophistication of the equations describing the be-

haviour of those elements. Increasing the number of elements and using more comprehensive equations are useful methods to increase accuracy, but there is a definite tradeoff between complexity and simulation speed. A modern IC design might use 100 to 200 transistors to implement a single functional block, compared to a handful of devices in a more traditional discrete design. If every device in a 200 transistor circuit is modelled with full complexity, simulation will be unnecessarily time-consuming.

The basis of a model is really the origin of the element topology and descriptive equations. Models may be numerically-based or physically-based. In a numerically-based model, element topology and behaviour are chosen to give the best fit to measured data at the device terminals. Noise behaviour of low-noise devices is frequently modelled in this way because such a device is typically used over a narrow range of operating conditions (those which give minimum noise figure). A physically-based model contains elements whose topology and behaviour are predicted from physical laws using a knowledge of device geometry and material properties. The advantages of a physically-based model are that the useful range of operating conditions may be larger than for a numerically-based model, and that models for devices with the same material properties and different geometry are more easily derived. This last idea is very important in IC design, where extensive use is made of transistors precisely scaled up from some known minimum geometry. A further advantage is that statistical variations of material properties and geometries can be treated directly within the model. Statistical design is becoming more important for high-volume products where robust, well-centered designs are required.

Common sense dictates that models for computer simulation should be accurate enough to describe the circuit at the level of detail desired, simple enough to require reasonable simulation run-times, and preferably physically-based in the sense that all model parameters are measurable physical quantities or are derived from such quantities. It is more difficult to decide what should be modelled in detail, and how, for a given circuit

under given conditions. To make these decisions, we need to know some details of the device physics and choose a reasonable mathematical approach to element description.

3.2 SPICE Models

All simulation for this thesis was performed using a commercial SPICE (Simulation Program with Integrated Circuit Emphasis) simulator [36]. The main reason for choosing SPICE is that it has become the de facto international standard for integrated circuit design due to the very wide distribution of the Berkeley 2G6 version. A second and related reason is that process vendors invariably provide SPICE models of their devices for customer simulation. SPICE models are physically-based, so they provide a convenient frame of reference for device physics pertaining to linearity. The models for passive components will first be discussed. More complex models for bipolar transistors will be introduced later.

In many IC processes the devices are isolated from each other by reverse-biased p-n junctions. These junctions introduce parasitic nonlinear capacitance shunting the devices. This implies signal degradation which varies with both amplitude and frequency. The mathematical treatment of this capacitance is discussed with transistor modelling later in this chapter.

3.2.1 Diffused Resistor Model

Figure 3-1 : Cross-section view of a diffused resistor

In Figure 3-1, a side-section view of a typical diffused resistor is shown. The resistor terminals are nodes 1 and 2; the metal contacts the P+ diffusion which forms the body of the resistor. The EPI contact is an ohmic contact to the n-type epitaxial layer, and is used to reverse bias the p-n junction formed by the epi material and the body of the resistor. This ensures that current will not flow across the junction but will remain confined to the body of the resistor. To achieve this, the epi contact is connected to the most positive dc voltage available (usually the positive power supply). The p-type substrate is common to every device on the chip, and is biased to the most negative dc voltage available (usually the negative power supply). In three dimensions, a guard ring is formed around the diffused resistor by the reverse-biased p+-n_{epi} region.

Figure 3-2 : Model schematic for diffused resistor

The resistor clearly has junction capacitance distributed along its length, and this is reflected in the model shown in Figure 3-2. The resistors are ideal elements modelled with the correct linear and quadratic temperature coefficients. To keep the element description simple and simulation time short, the total junction capacitance is distributed equally onto two internal nodes. Greater accuracy can be obtained by modelling the resistor as a lossy distributed structure.

So long as the P+ doping of the resistor body has a much higher concentration than the N doping of the epi layer, the depletion region of the isolation junction will be almost

entirely in the epi. But if the resistor body is either very thin or has a low dopant concentration (both conditions are used to make high-value resistors), the depletion region will make a significant incursion into the resistor body and the resistance will be raised. This implies a nonlinear $I(V)$ characteristic similar to that of a junction field-effect transistor (JFET) as shown in Figure 3-3.

Figure 3-3 : Diffused resistor carrying dc current showing (a) incursion of depletion layer into resistor body and (b) depletion layer widening

At terminal 2, the reverse bias is greater than at terminal 1 (the epi layer being connected to the positive power supply). The resistance per unit length increases and the capacitance per unit length decreases as one moves from 1 to 2 along the resistor. Both these effects are caused by the bias- and signal-dependent depletion layer width at the $P+/N$ junction. In situations where accuracy is extremely important, such a resistor would be modelled as a p-channel JFET where the epi contact is the gate and the resistor terminals are the source and drain. Note that the epi-substrate junction is dc-biased on both sides and is not subjected to signal swings.

3.2.2 MOS Capacitor Model

Another element usually provided on bipolar IC processes is some type of capacitor. One common type, the metal-oxide-silicon (MOS) capacitor, is shown in Figure 3-4.

Figure 3-4 : Cross-section view of a MOS capacitor

The capacitor plates are the metal plates labelled MP and the P+ region contacted by PP. As with the resistor, an EPI contact is used to reverse bias the junction between the P+ lower plate and the epi material, thus isolating the device. Once again the substrate is connected to the negative power supply.

Figure 3-5 : Model schematic for MOS capacitor

The model for such a device is shown in Figure 3-5. The MP/PP capacitor dielectric is the field oxide of the process, so this capacitor is nearly ideal and very linear. The parasitic capacitance between PP and the epi is usually minimized by returning EPI to the positive power supply. Sometimes, however, maximum capacitance per unit area is desirable and the nonlinearity of a junction capacitor is not too objectionable. In this case EPI is connected to MP to form one plate of the capacitor, with PP forming the other plate. Under these conditions, the voltage on EPI/MP must always be more positive than the voltage

on PP, or the epi/P+ diode will be forward-biased.

3.2.3 Bipolar Transistor Model

Resistors and capacitors are the most common passive components provided in IC processes, and they exhibit fairly simple nonlinear behaviour. But the bandwidth and linearity of a circuit are determined much more strongly by the performance of the transistors, especially those in the signal path. The nonlinearities associated with transistors include not only the $C(V)$ relationship for a reverse biased junction, but many other effects of varying subtlety. After a brief explanation of transistor operation, I will introduce the basic Gummel-Poon model and discuss each nonlinearity.

A modern planar NPN transistor is shown in Figure 3-6. The emitter contact is labelled E, the two base contacts are labelled B and the collector is labelled C. The normal operating mode (or forward active region) for such a device is with the emitter-base junction forward biased and the collector-base junction reverse biased. Under these conditions, electrons are injected from the emitter junction into the base, where they diffuse across the very small distance to the collector junction. Here they are swept up by the large drift field in the collector-base depletion layer. The N^+ deep collector diffusion and the N^+ buried layer are used to provide a low-resistance path from the active collector region under the base and emitter to the contact on the surface of the chip. As always, the substrate is connected to the negative power supply, and the collector-substrate junction must always be reverse biased.

$$I_C = I_S \left(\exp \left(\frac{V_{BE}}{V_T} \right) - 1 \right) \quad (3.1)$$

and capacitance versus voltage for reverse bias:

$$C_J = C_{J0} \left(1 + \frac{V_R}{\psi_j} \right)^{-m} \quad (3.2)$$

where I_C is collector current, I_S is the transport saturation current, V_{BE} is base-emitter voltage, V_T is the thermal voltage equal to kT/q , C_J is junction capacitance, C_{J0} is the zero-bias value of junction capacitance, V_R is the reverse voltage across the junction, ψ_j is a barrier voltage which depends on doping and temperature and m is the junction grading coefficient.

The split of collector-base capacitance using elements CBC and CBCX is an attempt to model the distributed nature of the physical collector-base capacitance as shown in Figure 3-8. A gradual lateral voltage drop occurs due to base current flow in the bulk base material. At the emitter-base junction, this lateral voltage drop causes the forward bias on the junction to be lower at the center of the emitter than at the edges. This condition is called current crowding [40], and it causes most of the carrier injection to occur at the emitter periphery. At the collector-base junction, the reverse bias is greater at the center of the emitter than at the edges. Greater reverse bias implies a wider depletion region and lower capacitance, so capacitance per unit area will be lower near the center of the emitter than at the periphery.

The great advantage of the GP2 model over the older Ebers-Moll (EM) model is that it attempts to account for high-level injection effects and basewidth modulation effects in a self-consistent manner. These effects are the degradation of current gain and transit time at high collector currents, the Early effect and the current dependence of parasitic

base resistance.

Figure 3-8 : Distribution of collector-base capacitance across base resistance for the planar NPN transistor

The rigour of the model derives from its physical basis whereby charge concentrations in various regions of the device control the behaviour of the model elements. The value of the model to the circuit designer is that it accurately predicts a wide range of phenomena under useful device operating conditions. Because of the change in device geometries since the introduction of the GP model, some of the model parameters now have the status of fitting variables rather than purely physical quantities.

High-level injection occurs when the concentration of injected minority carriers in the base is comparable to the majority carrier concentration (approximately the base doping in the simplest case). Since space-charge neutrality is maintained in the base, the majority carrier concentration increases by the same amount as the minority carrier concentration, and the exponential dependence of collector current on base-emitter voltage weakens to:

$$I_{CH} \propto \exp\left(\frac{V_{BE}}{2v_i}\right) \quad (3.3)$$

This phenomenon is known as the Webster effect [41], and it causes a fall-off or

"knee" in the current gain versus collector current characteristic of the transistor.

Transit time degradation at high current levels is due to reduction of the aiding field at the collector base junction and to the Kirk effect [42]. When the concentration of electrons being swept across the collector-base junction approaches the collector epitaxial doping concentration, the electrons in transit form a negative space charge which partially cancels the positive space charge due to the ionized donors on the collector side of the depletion layer. This reduces the aiding field seen by the electrons arriving from the base. In the extreme case where the concentration of arriving electrons approximately equals the background epitaxial doping, space charge neutrality is achieved for a thin region of collector near the metallurgical junction. The width of this neutral region adds to the effective basewidth, increasing transit time τ_f . It is this base "widening" that is referred to as the Kirk effect. To generalize grossly, the Webster effect introduces amplitude distortion and the Kirk effect introduces phase distortion. The Kirk Effect is the dominant high-level injection effect in modern bipolar IC transistors.

The basewidth modulation effects arise from the bias-dependent incursion of the emitter and collector depletion layers into the base. The most noticeable effect of basewidth modulation is the Early effect [43], where a widening collector-base depletion layer (caused by increasing collector-base reverse bias) narrows the base, causing an increase in diffusion current. The result is that the transistor is no longer an ideal transconductance (voltage-controlled current source), but has a finite output impedance. For hand calculation, Equation (3.1) is modified to:

$$I_C = I_S \left(\exp \left(\frac{V_{BE}}{V_T} \right) - 1 \right) \left(1 + \frac{V_{CE}}{V_A} \right) \quad (3.4)$$

where V_A is the Early voltage extrapolated from a plot of I_C versus V_{CE} as shown in Figure 3-9. Note that for a wide range of base (or collector) currents, the extrapolated Early volt-

age is constant.

Figure 3-9 : Extrapolation of the common-emitter output characteristics to determine the Early voltage

The bulk base resistance of a transistor is also affected by junction bias voltages. Unlike the emitter and collector bulk resistances, the resistance of the active base region is that of a thin sheet of material between two diode junctions. The importance of this resistor to dc current crowding has already been discussed, but it also has an important effect on frequency response, noise behaviour and linearity. For this reason, the GP2 model includes a base resistor which varies with both current injection level and bias voltage.

Having discussed some of the effects covered by the GP2 model, it is important to consider some effects that are not modelled. One of the most important of these is breakdown; that is, the avalanche multiplication of thermally-generated carriers in the collector-base depletion region. In large-signal terms, this corresponds to a sudden decrease in output impedance when the collector-base reverse bias exceeds the breakdown voltage BV_{CBO} (measured with the emitter open circuit). The avalanche multiplication factor M is defined as [44]:

$$M = \left(1 - \left(\frac{V_{CB}}{BV_{CBO}} \right)^n \right)^{-1} \quad (3.5)$$

where n is a parameter defining the sharpness of the breakdown. Breakdown can be modelled using a current-controlled current source connected between collector and base as shown in Figure 3-10.

Figure 3-10 : NPN model schematic modified to simulate collector-base breakdown

The main problem with this approach is its effect on convergence during numerical solution. M in Equation (3.5) increases very rapidly near breakdown, and the circuit will fail to converge if an iteration carries V_{CB} beyond BV_{CBO} . Some commercial versions of SPICE have convergence enhancements which constrain the iteration range, but usually breakdown is left unmodelled. Of course, this means the designer must pay attention to the collector base bias for every device in the circuit under all operating conditions. Some simulation packages have electrical rule checker (ERC) features which perform this audit automatically for a variety of operating conditions. A report is produced which details breakdown conditions, saturation, high-level injection and so on for every device in the

circuit.

Another unmodelled effect can be described as dynamic temperature variation. For the most part, SPICE pays careful attention to temperature in its bias analysis. The parameters XTI, EG and XTB are used to control the values of IS, BF, BR, ISE, ISC, CJE, VJE, CJC and VJC at temperatures different from the nominal temperature (usually 27 deg C) where the model parameters are extracted. This parameter correction is performed only once, during the operating point analysis. As a result, SPICE assumes that device temperature remains constant regardless of signal-induced deviations from the initial operating point. It also assumes that the parasitic bulk resistances RB, RE and RC are independent of temperature.

Real circuit behaviour is quite different. As large signal swings appear across the device output, the instantaneous power dissipation changes and so does the temperature. The bulk resistances have significant temperature coefficients (typically 1500 ppm/deg C for RB and 1400 ppm/deg C for RE [45]). More importantly, all properties which are explicitly functions of temperature change value. For example, base-emitter voltage decreases with increasing temperature (at constant current), while current gain increases. Dynamic temperature variation can be due to device self-heating, or to cross-heating by an adjacent device. It is possible to model some of these effects; the means for doing so are discussed in the next chapter. Parameter variation with instantaneous temperature is not included in the basic GP2 model.

At sufficiently high frequencies (near $f_T/3$), the GP2 model is inadequate in its representation of parasitics, especially the intrinsic/extrinsic division of the base and collector resistances. Figure 3-11 shows the model extension for an NPN transistor, with the original GP2 core inside the dashed lines. Now the collector-substrate junction capacitance "splits" the intrinsic and extrinsic (RCX) portions of RC, and the base-charge-dependent intrinsic RB is split from the relatively constant RBX component. The additional distribu-

tion of collector-base junction capacitance also provides more accuracy at high frequency. The price paid for this accuracy is two more nodes per transistor, increasing simulation time.

Figure 3-11 : Extended Gummel-Poon model derived from basic GP2 model

Further refinements to device modelling will be introduced where appropriate in Chapter 4.

3.3 Simulation Using SPICE

At this point, attention shifts from modelling considerations to those of simulation. In some ways, the choice of simulation method is less important than the choice of model because any worthwhile simulation must yield results which agree with measurements of the physical circuit. The introduction to this thesis briefly compared frequency-domain, time-domain and hybrid simulation methods. I selected the SPICE simulator (Simulation Program with Integrated Circuit Emphasis) for three reasons.

First, SPICE has become the de facto standard circuit simulation program world-wide. From the original Berkeley "public distribution" version (2G6), commercial variants have been driven with enhanced convergence capabilities, superior graphics and more ca-

pable analysis modes.

Second, many vendors perform iterative device characterizations based on a cycle of device measurement, SPICE model modification and SPICE simulation, so as to enhance agreement between simulated and measured results. If the models are translated to another topology or set of describing equations, there is a risk that the subsequent simulation will not agree with measured results because different assumptions were used in the model derivation.

Third, SPICE has three useful ways to simulate linearity, as well as a comprehensive noise analysis. This is important because of the many different ways in which distortion is specified, and because linearity compensation often involves a degradation of noise figure.

The version of SPICE used for this work, Tekspice [46], features enhanced convergence, superior graphics capabilities, nested variable sweeping and a full range of analyses: bias, dc, ac, transient and Volterra. The last three are directly useful in distortion analysis.

3.3.1 Volterra Analysis

In the Volterra analysis [47], a simplified nonlinear model is generated from the Ebers-Moll or Gummel-Poon SPICE parameters as shown in Figure 3-12a. Three nonlinearly controlled sources i_1 , i_2 and i_3 are added as shown in Figure 3-12b to account for exponential junction nonlinearities, junction capacitances and the Early effect.

Figure 3.12 : Nonlinear (a) and quasilinear (b) models for Volterra analysis

Although the control equations for the sources are nonlinear, a perturbation method can be used to generate a solution by nodal analysis. Knowledge of the interreciprocal properties of the adjoint network [48] allows the contribution of each nonlinear source in each device to the overall distortion to be calculated. This very useful property is shared by the noise analysis.

The Volterra analysis has the advantages of elegance and speed. Its disadvantages are twofold. First, the Ebers-Moll model is less comprehensive and accurate than the Gum-

mel-Poon model (and the charge control mechanism is too intractable to factor into the quasilinear model of Figure 3-12b). Second, the Volterra series expansion method itself begins to lose accuracy when signal current magnitudes exceed 5 to 10 per cent of their bias values. This problem occurs because the Volterra series is usually truncated at three terms due to the difficulty of identifying and expanding higher-order Volterra kernels. For weakly driven class A amplifiers, the "individual distortion contributor" information is very valuable, particularly when used with the individual noise contributor feature of the noise analysis. This is the most common method used to optimize dynamic range. The circuits considered for this thesis have signal currents 50 per cent or more of their bias values, so other methods are necessary for more accuracy.

3.3.2 Differential Gain Analysis

One method which is quite accurate and provides results very quickly is the differential gain analysis. The principle of this analysis is simple: a series of ac small-signal analyses are performed over a range of dc values which corresponds to the peak-to-peak swing of the input signal. Conceptually, it is similar to evaluating the local or differential slope at a series of points along a curve representing the input signal as a function of time. Since the gain magnitude and phase (or any other small-signal parameter) can be determined at each point, this analysis is most useful for looking at magnitude and phase compression in response to a sinusoidal or other "smooth" input. For example, suppose an amplifier's input is biased to -2 V, and is excited by an input sinusoid of 200 mV peak-to-peak. The dc input bias voltage would be swept from -2.1 V to -1.9 V in a number of equal steps, and at each step an ac analysis over some frequency band can be performed (normally only one frequency is chosen). When these "nested" analyses are complete, the gain magnitude or phase can be plotted against the value of dc input as a sort of compression plot. If the circuit is perfectly linear, the plot will be a flat line. As noted in chapter 2, most real circuits experience compression (reduced gain magnitude at the signal extremes). One important restriction with this method is that the input must be dc-coupled to the output

to get realistic results. In ac-coupled circuits the coupling capacitors must be replaced by dc voltage sources whose values are determined from a quiescent bias analysis. Figure 3-13 shows how this is done.

The differential gain analysis does not exactly simulate the compression behaviour of circuits because of its quasistatic nature. However, the compression predicted for amplitude and phase is very similar to the measured distortion for continuous sinusoidal excitation. This method of analysis does account for high-level injection and current-dependent base transit time, as well as some other base charge controlled effects. Perhaps most usefully, differential gain compression simulation can be performed in conjunction with dynamic thermal variation due to device self-heating or cross-heating. This is discussed in more detail in Chapter 4.

Figure 3-13 : Replacement of coupling capacitors for differential gain analysis

Determine $V_{b1} \dots V_{b4}$ from bias analysis

3.3.3 Transient Analysis

The third distortion analysis method can be used to model compression, but is more often used to model intermodulation or cross-modulation. It is the very direct method of performing a full transient analysis for the circuit with the desired excitation, followed by a Fourier transform. Transient or time-domain analysis is the most accurate of the methods because the nonlinear differential equations are formulated and solved directly, rather than introducing quasilinear models which necessarily make additional ap-

proximations. Transient analysis is also best able to handle inputs which are arbitrary functions of time. Compression behaviour can be easily examined in time or frequency domain, but the transient analysis model is most useful to study intermodulation performance. It is also indispensable for testing slew rate performance.

When SPICE has completed a transient analysis, a fast Fourier transform (FFT) can be used to display the results in the frequency domain. The number of points transformed, N , must be an integer power of two ($N = 2, 4, 8, 16, \dots$). The time duration to be transformed is T . The function values are defined at equally-spaced intervals from $t=0$ to $t=(N-1)\Delta t$. In the frequency domain, the two-sided bandwidth of the transform is B , centered on $f=0$. The resultant frequency points are spaced a distance Δf apart from $f=-B/2$ to $f=((N-1)/N)(B/2)$. The intervals Δt and Δf represent the time-domain and frequency-domain resolutions, respectively. Figure 3-14 makes clear the relationships between N , Δt , T , Δf and B . Note especially that the frequency resolution is the reciprocal of the time duration. This means that closely-spaced test frequencies in a two-tone intermodulation test will demand a long time duration if the tones are to be resolved.

Figure 3-14 : Frequency and time domains with relations between FFT parameters

The spectrum calculated by the FFT has a "folding" property: spectral replicas centered at $f = \pm nB$ fold down to $f = 0$. To prevent aliasing [49], the frequency response must either inherently roll off or be filtered so as to have negligible spectral content beyond $B/2$. For a wideband circuit, large B implies small Δt in the transient analysis, again increasing the simulation time.

The most useful results are obtained when the excitation frequencies (and therefore their harmonics and intermodulation products) fall directly onto the points in the FFT. If they do not, the output frequencies will fall between points in the FFT as shown in Figure 3-15a, and the magnitude of the output tone cannot be easily determined. This condition is known as scallop loss [50]. Correct choice of excitation frequency and transform parameters is shown in Figure 3-15b, where the amplitude of the tone can be determined directly.

Figure 3-15 : Incorrect (a) and correct (b) choice of excitation frequency and transform points for FFT

The shape $W(f)$ in Figure 3-15 represents the frequency-domain representation of the windowing function. The purpose of the windowing function is to ensure that the time series submitted to the Fourier transform tapers to zero at the start and end points. The main issues associated with windowing are coherent gain, main lobe width and maximum sidelobe level [51]. Coherent gain refers to the relative magnitudes of the windowed and unwindowed time domain functions. Coherent gain is constant for all tones within the range of the transform, and is therefore not objectionable: it can be corrected by scaling the spectrum with a constant. Many commercial FFT implementations perform this scaling automatically. Main lobe width of the transform of the window determines the frequency resolution achievable. If two or more tones fall within one frequency bin Δf , the tones will not be resolved, but will be smeared together on a power basis. Finally, sidelobe levels determine the amplitude resolution of the transform. The situation may arise where the main

lobe of the window function "detects" a very low-level distortion product, but a sidelobe of the window detects a very high-level signal. this situation is shown in Figure 3-16, and will result in an unintended sidelobe contribution to the simulated level of the low-level product.

Figure 3-16 : Undesired contribution to measured distortion level by window sidelobe detection of high-level signal

The Blackman-Harris window [52] is used for all simulations in this thesis. Coherent gain for this window is 0.36, the sidelobe level is -92 dBc and the 6 dB bandwidth is $2.72((N-1)(B/N))$. Careful use of the differential gain analysis and the transient/FFT analysis yielded simulation results which accurately predicted the behaviour of the fabricated ICs. The 0.1 dB and 1 degree compression points were predicted to within 1 dB in level, and intercept points were accurate to 3 dB for the types of circuits considered in this thesis.

3.4 Guidelines for IC Design Using CAD

Some brief commentary on the design process is now appropriate, because an intelligent application of these techniques can save much time and trouble in the design of integrated circuits. The designer begins with the specifications and one or more trial topol-

ologies for the circuit. Each topology consists of the signal path components (including both forward and feedback elements) along with voltage and current bias sources. Each circuit topology is tried first with the signal path components modelled by real devices, and the bias sources modelled as ideal voltage and current sources. This allows quick selection of device sizes and operating points and greatly speeds simulation. At this stage the circuit should have comfortable margins on all performance specifications, because degradations will be caused when the ideal elements are replaced by real devices. When the topology, device sizes and operating points have been chosen, the ideal sources can be gradually replaced by real devices. The reason for doing this gradually is to develop a picture of which devices contribute most to distortion and noise. Once the process is complete and all devices are correctly modelled, operation over process and temperature extremes should be checked. Finally, parasitics should be added and a comprehensive round of simulation should be performed. If worst-case distortion margins are less than ten per cent, the designer should examine the effects of thermal distortion and device mismatches.

Beyond these basic guidelines, it is difficult to generalize further about the design process for ultralinear ICs. There is a body of expertise touching on device physics, circuit design and thermal design which provides ideas for robust design, and also exposes the constraints on linearity which must be understood by any designer who seeks maximum performance. This collection of ideas is the subject of the next chapter.

4 Design Considerations

The design of analog circuits is notoriously difficult to define in a closed-form algorithmic manner. The design considerations presented here are not a methodology but a set of constraints on achievable performance. The order of presentation is roughly from most straightforward to most subtle.

4.1 Device Design and Operating Point

Abraham and Meyer [19] studied the issue of device design for optimized distortion performance. Their procedure was to use a fairly simple model for the transistor and to write explicit equations for gain, third-order intermodulation (IM3) and second-order intermodulation (IM2). Examining these equations and drawing conclusions about the physical device parameters, they found that epitaxial layer thickness and resistivity have a very strong effect on the distortion performance of bipolar transistors. To minimize Kirk effect and parasitic bulk resistance, the epi material should be thinner and more heavily doped, while to achieve high BV_{CEO} and high gain the epi layer should be thicker and more lightly doped. In addition to this conflict, they found that nonlinearity due to the bias-dependent collector-base junction capacitance could be greatly reduced by completely depleting the epi layer below the base (see Figure 4-1). The collector-base reverse bias needed to produce this depletion is the depletion voltage V_D , and it is generally useful for this voltage to have a low value (certainly lower than the breakdown voltage BV_{CBO}). This implies a thin, lightly-doped epi layer.

Figure 4-1 : Full depletion of collector epi-layer

It is very useful to know the value of the depletion voltage for a given process, and it is easily calculated:

$$V_D = \frac{t^2 q N_D}{2\epsilon_s} - \psi_{jc} \quad (4.1)$$

where t is the epi layer thickness below the base, q is the electronic charge, N_D is the epi layer doping concentration, ϵ_s is the permittivity of silicon and ψ_{jc} is the collector-base barrier voltage. For the Tektronix SH3 process used in this work, $t = 0.8 \mu\text{m}$, $N_D = 2.5 \times 10^{16} / \text{cm}^3$, $\epsilon_s = 11.8 \times 8.854 \times 10^{-14} \text{ F/cm}$ and $\psi_{jc} = 0.65 \text{ V}$; from Equation (4.1) $V_D = 11.6 \text{ V}$. Since the minimum value of BV_{CBO} is 15 V , the maximum output signal swing for “depleted collector” mode is less than 4 V . The supply headroom required may be prohibitive for low-power applications, and it is important to remember that the GP2 model does not account for collector depletion. Devices which see large collector-base voltage swings should be biased with the highest collector-base reverse bias possible (subject to breakdown, dissipation and supply headroom limitations). Figure 4-2 shows transconductance magnitude linearity and phase linearity as functions of collector-base reverse bias for several values of base-emitter voltage. As noted in Chapter 3, breakdown is not normally modelled so the designer must consider voltage extrema on each relevant node individually.

Figure 4-2 : Magnitude and phase of transconductance as functions of collector-base voltage at several values of emitter-base voltage (simulated)

Another operating point consideration is the need to ensure that the peak collector current remains well below the knee current for high-level injection. Exceeding the knee current will usually introduce pronounced magnitude and phase distortion into the output current of a BJT, especially if the device is being used as a transconductance. Figure 4-3 shows the transconductance magnitude linearity and phase linearity as functions of base-emitter voltage for several values of collector-base bias. In this case the device is designed for peak f_T at 12 mA collector current, and the extrapolated knee current (IKF) is about 80 mA, while the high-level threshold current for base transit time (ITF) has an extrapolated value of about 180 mA. Of course, these values of IKF and ITF are fitting pa-

rameters, not physical currents.

Figure 4-3 : Magnitude and phase of transconductance as functions of base-emitter voltage at several values of collector-base voltage (simulated)

It is instructive to consider the effect of parasitic base and emitter resistances on transconductance. In Figure 4-4a, an ideal transistor is shown with the implicit relationship between collector current and base-emitter voltage:

$$I_C = I_S \exp\left(\frac{V}{V_T}\right) \quad (4.2)$$

Figure 4-4 : Transconductance degradation by parasitic base and emitter resistance

When this device is degraded by parasitic base and emitter resistances R_B and R_E as shown in Figure 4-4b, the collector current expression becomes:

$$I_{CX} = I_S \exp\left(\frac{V - I_B R_B - I_E R_E}{v_t}\right) = \frac{I_C}{\exp\left(\frac{I_C}{\beta v_t} R_{eq}\right)} \quad (4.3)$$

where $R_{eq} = R_B + (\beta+1)R_E$. The transconductance is modified by a factor which depends directly on collector current, and also indirectly through β and the R_B component of R_{eq} . This shows clearly that parasitic resistance not only causes increased loss and noise, but also an increase in distortion.

The product of parasitic base resistance and collector-base junction capacitance is a useful indicator of high-frequency transistor performance. The $R_B C_{jc}$ product is roughly constant for a given process, regardless of device scaling. This is because R_B scales inversely with device size, while C_{jc} scales directly. As a rule of thumb, analog bipolar processes should have $R_B C_{jc}$ less than 1/3 the value of τ_f if noise figure and intermodulation are important specifications.

If the impedance seen by the collector is non-zero, the collector-base and collector-substrate capacitances become important. In particular, the distribution of these capaci-

ances over the base and collector parasitic resistances must be modelled accurately, especially in high slew rate circuits with large gain. For this reason, the basic GP2 model (Figure 4-5a) is modified to the extended-frequency GP2 model in Figure 4-5b. R_{BX} and R_{CX} are the extrinsic components of R_B and R_C respectively. The simulated performance of the extended-frequency model is much closer to measured data at frequencies above $f_T/5$, but each device adds two nodes and one extra nonlinear element to the circuit compared to the GP2 model.

Figure 4-5 : Conventional (a) and extended-frequency (b) SPICE GP2 models

Another design consideration arises from the data provided by most vendors for f_T as a function of operating conditions; usually I_C and V_{CE} . An example of such data is given in Figure 4-6.

Figure 4-6 : f_T versus I_C and V_{CE} for an E60X NPN transistor (simulated)

The data is useful for the design of current-mode circuits, and less so for voltage-mode circuits where the collector load impedance is substantially greater than zero. Usually the designer selects a bias current significantly less than the current for peak f_T . This is done to allow for signal swings and temperature variations; in all cases the peak collector current must be prevented from entering the high-level injection region of the curves. Operating below the current for peak f_T will also reduce shot noise.

4.2 Single Transistor Stages

Any circuit designer will be familiar with the small-signal properties of the three basic single-transistor stages: common emitter (CE), common base (CB) and common collector (CC). Their noise properties are likewise familiar to most RF engineers, but it is more interesting to compare their relative merits from the standpoint of wideband linearity.

A comparison of CE, CB and CC circuits was made under identical bias conditions. The devices were configured as approximately 200 ohm sources driving 200 ohm loads, and each was driven with the appropriate signal (voltage or current source). The circuits are shown in Figure 4-7, and the gain magnitude linearity and phase linearity as functions

of input level are studied. This type of information is directly useful to the designer with a fixed power budget who must drive a matched load off-chip.

Figure 4-7 : Single-transistor gain stages used for compression comparison

The E60X device selected was designed for a peak f_T of 6 GHz at a collector current of 12 mA. The bias conditions chosen were ($V_{CB} = 3$ V, $V_{BE} = 0.85$ V), which gives an emitter current of about 10 mA. The differential gain test described in Chapter 3 is used to determine the gain magnitude compression under large-signal conditions. Since the circuit must be dc-coupled, a bias analysis is performed in each case with the load resistor absent and the quiescent output voltage found with this analysis is used to set the dc value of the voltage source "voq" so that no load current flows under quiescent conditions. The output powers for each circuit at 0.1 dB compression are described in Table 4.1:

Table 4.1: Output 0.1 dB Compression Power for Single-tansistor Stages

Type	$CP(0.1\text{ dB})_{\text{out}}$
CE	-10.1 dBm
CB	+3.4 dBm
CC	+7.4 dBm

These results strongly imply that CB and CC stages have a large advantage over CE stages if wideband linearity and fixed power dissipation are important, as is usually the case in IC design. Because of its superior noise properties, the CE stage is indispensable for front-end design, but in such a situation it must be used with care if the receiver is to handle large spurious signals without distortion. The degraded linearity of the CE stage is due to the large (nonlinear) Miller capacitance seen at the base node as indicated in Figure 4-8.

Figure 4-8 : Equivalent Miller capacitances for a voltage amplifier with shunt capacitive feedback

The input capacitance to ground is effectively the collector-base capacitance multiplied by the voltage gain of the amplifier. The linearity of the CE stage can be greatly improved with the addition of series or shunt resistive feedback, but this degrades noise performance and reduces gain.

4.3 Effects of Impedance

In any wideband system, high-frequency signals will be routed using "lossless" transmission lines if the routing distance is significant compared to the minimum wavelength used. Such lines are completely defined by their inductance per unit length (l) and their capacitance per unit length (c):

$$v_p = \lambda_g f = \frac{1}{\sqrt{lc}} \quad (4.4)$$

$$Z_0 = \sqrt{\frac{l}{c}} \quad (4.5)$$

where v_p is the velocity of propagation, λ_g is the "guide wavelength", f is frequency and Z_0 is the characteristic impedance. Power transfer efficiency is maximized and linear distortion is minimized if the source and load impedances at each end of a transmission line are matched to the characteristic impedance of the line.

The choice of Z_0 for a particular application depends on several factors. Interconnections to other equipment or test fixtures are important considerations; these are normally at 50 ohms or 75 ohms. For transmission lines printed on a substrate, the maximum value of Z_0 is determined by the minimum metal linewidth allowed in the process, and by the effect of parasitics on magnitude flatness and phase linearity. If printed filters, couplers or matching structures are to be used, they may also determine Z_0 .

The choice of Z_0 affects the linearity requirements for wideband integrated circuits. Many such circuits have high-impedance inputs and low-impedance outputs as shown in Figure 4-9a. The simplest way to match them is by shunting the input with a resistor of value Z_0 and placing a resistor of the same value in series with the output as shown in Figure 4-9b.

Figure 4-9 : Source and load matching for a typical wideband IC

If the input voltage is V as shown, the output voltage across the load is $KV/2$ and the output power delivered to the load is $P_{out} = (KV)^2 / (4Z_0)$. To achieve this, the low-impedance output must be capable of driving a voltage $2V_{out}$ and a current of V_{out}/Z_0 without appreciable distortion. This requirement sets absolute constraints on the output stage design, but there is also obviously a requirement on the input of the circuit: the input stage of the amplifier must not distort appreciably for input signal amplitudes less than or equal to the maximum output voltage divided by the gain of the amplifier.

Consider the circuit of Figure 4-10: a CE amplifier with series emitter degeneration followed by a CC output stage.

Figure 4-10 : Amplifier test circuit

Such a circuit has high input impedance and low output impedance, so the output is series-matched using R4; R5 is the load resistor. If V_{iq} (the input quiescent voltage) and resistors R1, R2 and R3 are kept constant, and R4 and R5 are set equal to a value Z which is varied over the range of 25 ohms to 200 ohms, the output 1 dB compression point as a function of Z is shown in Figure 4-11. The domain $Z < Z_{opt}$ (region A) is called the output-limited region; here the output is incapable of sourcing or sinking enough current at the extremes of the signal swing. Region B, where $Z > Z_{opt}$, is the input-limited region; here the high input voltage drives the input stage beyond its linear region.

Figure 4-11 : Output compression point versus impedance level for the circuit of

Figure 4-10

For a given circuit, Figure 4-11 implies that there is an optimum range of impedances to maximize output compression point, with all other performance criteria such as gain, power consumption and noise figure held constant. Conversely, where impedances are defined by system or subsystem considerations, the circuit can be designed so that the optimum impedance Z_{opt} coincides with the system characteristic impedance Z_0 . For a specific circuit topology and output compression point, a lower value of Z_0 will demand

higher bias currents and therefore greater power consumption and shot noise. It is therefore important to select system impedance levels carefully, especially where access to thin-film substrates and multichip module (MCM) technology makes a choice of higher Z_0 possible.

It is often convenient to use lumped-element or distributed impedance transformers to raise the impedance at the input or output of a chip. This technique is especially effective when the transformers are used to convert single-ended signals to differential ones which can be processed more efficiently with an integrated circuit. The use of impedance transforming baluns is illustrated in Figure 4-12.

Figure 4-12 : Impedance matching of a wideband IC using balun transformers

V_{bias} is the dc bias voltage required at the input of the amplifier; the input is 1:1 shunt matched and the output is 4:1 series matched. If Z_0 is 50 ohms, the chip outputs drive a total resistance of 400 ohms (100 ohms per side series matching resistance plus the 200 ohm transformed load impedance). The drawback to using transformers is that they limit bandwidth at both low and high frequencies.

4.4 Differential Versus Single-Ended Circuits

The advantages of differential over single-ended circuits have been recognized since the earliest days of integrated circuit design. These include increased immunity to

power supply noise and other sources of interference, a greater variety of useful design techniques and usually improved wideband linearity.

It is easy to illustrate how the use of a differential signal path can improve linearity. Consider the circuit of Figure 4-13. The amplifiers are identical voltage-in voltage-out devices with a memoryless nonlinear characteristic:

$$V_y = a_1 v_x + a_2 v_x^2 + a_3 v_x^3 + \dots \quad (4.6)$$

Figure 4-13 : Identical imperfect amplifiers operated in differential mode

If the input voltage is $v_{in}(t) = V_{in} \cos(\omega t)$ and the input transformer is perfect, the inputs to the amplifiers will be 180 degrees out of phase with each other and have amplitude $V_{in}/2$:

$$V_1(t) = \left(\frac{V_{in}}{2} \right) \cos(\omega t) \quad (4.7)$$

$$V_2(t) = -\left(\frac{V_{in}}{2} \right) \cos(\omega t) \quad (4.8)$$

When Equations (4.7) and (4.8) are substituted into Equation (4.6), we obtain:

$$V_3(t) = \left(a_1 \frac{V_{in}}{2}\right) \cos(\omega t) + \left(a_2 \frac{V_{in}^2}{4}\right) (\cos(\omega t))^2 + \dots \quad (4.9)$$

$$V_4(t) = \left(-a_1 \frac{V_{in}}{2}\right) \cos(\omega t) + \left(a_2 \frac{V_{in}^2}{4}\right) (\cos(\omega t))^2 + \dots \quad (4.10)$$

Finally, the output is just the difference between $v_3(t)$ and $v_4(t)$ if the output transformer is perfect:

$$V_{out} = (a_1 V_{in}) \cos(\omega t) + \left(a_3 \frac{V_{in}^3}{4}\right) (\cos(\omega t))^3 + \dots \quad (4.11)$$

First, note that the linear voltage gain is a_1 for both the differential and single-ended circuits. Second, note that all terms of even degree vanish, so even-order distortion is intrinsically eliminated from the output (again, this assumes perfect transformers and identical amplifiers). This property is very valuable in true wideband amplifiers with greater than one octave bandwidth, because in these cases even-order intermodulation products end up in-band. Finally, note that the coefficient of the $(\cos(\omega t))^3$ term is $1/4$ as large as for a single-ended amplifier with the same input amplitude. This reduction in distortion is even greater for odd-order terms of higher degree; for example, the coefficient of the $(\cos(\omega t))^5$ term is $1/16$ as large as for the single-ended amplifier.

Two issues that arise when comparing differential and single-ended circuits are device placement and power consumption. While extra devices are essentially free on an IC substrate, the placement of the devices in a differential circuit is very important because the circuit must have electrical and thermal symmetry. Use of electrical symmetry guarantees that the path lengths and parasitics will be the same in both "halves" of the circuit, while thermal symmetry ensures that "pairs" of devices will have identical properties. If a

differential circuit is to be compared fairly against a single-ended one, it is important to consider any difference in power consumption between the two. The usual topology choice is a differential signal path and a common-mode bias network; this gives the best noise performance and common-mode rejection.

Consider the circuits of Figure 4-14. The circuit of Figure 4-10 is considered as a reference block in three situations.

Figure 4-14 : Single-ended and differential amplifier test circuits

In Figure 4-14b, the compression of the single-ended circuit is evaluated with a series-matched 100 ohm load. In Figure 4-14c, the compression of the differential circuit with

a series-matched 200 ohm load is evaluated, while in Figure 4-14d the load is 100 ohms. The output 1 dB compression points for these three cases are:

Table 4.2: 1 dB Compression Point for Circuits of Figure 4-14

Circuit	CP(1 dB)
4-14b	-1.07 dBm
4-14c	+1.92 dBm
4-14d	-0.18 dBm

It is reasonable to conclude that the choice of a differential topology provides a moderate improvement in compression performance and a large improvement in intermodulation performance.

A further advantage of differential circuits is their ability to "double" effective slew rates. For example, suppose an amplifier must drive a capacitive load, and the required slew rate is $1000 \text{ V}/\mu\text{s}$. If a single-ended driver is used, the output must be able to achieve $1000 \text{ V}/\mu\text{s}$, but if the load can be driven differentially, the complementary outputs need only be able to slew at $500 \text{ V}/\mu\text{s}$ individually to provide the required slew rate, because they are ramping in opposite directions.

4.5 Effects of Nonideal Bias Circuits

Near the end of Chapter 3, it was pointed out that performance specifications are normally degraded if the ideal voltage and current sources used for initial design are replaced by real devices. This is because real devices have associated linear and nonlinear parasitics which attenuate and distort the signal as it passes through the circuit.

In the example cascode amplifier of Figure 4-15, the parasitics G_P and C_P for the current sink and series resistance R_S for the voltage source are shown. The effect of each

type of nonideal bias will be examined, then an estimate of the bias circuit contribution to overall distortion can be made.

Figure 4-15 : Cascode amplifier with non-ideal resistance R_S , conductance G_P and capacitance C_P

The effect of a resistance R_S in series with the cascode bias voltage source is to increase the input impedance of the common base stage. In Figure 4-16a, a transistor is biased at collector current I_C , with the collector at ac ground and the base tied to ac ground through a resistance R_B . R_B is the sum of the internal parasitic base resistance and the equivalent external bias resistance. In Figure 4-16b, the transistor has been replaced by its small-signal hybrid- π equivalent circuit and a test current i_x is applied at the emitter

2

PM-1 3 1/2"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

Figure 4-16 : Emitter impedance of a bipolar transistor

The node voltage v_2 is found using nodal analysis, and the impedance Z_E looking into the emitter is then $Z_E = v_2/i_x$. This is easily found to be:

$$Z_E = \frac{G + g_\pi + sC_\pi}{G(g_m + g_\pi + sC_\pi)} \quad (4.12)$$

where $G = 1/(R_S + r_b) = 1/R_B$, $g_\pi = 1/r_\pi = g_m/\beta$ and $g_m = I_C/v_t$. For low frequencies and $\beta \gg 1$, we find Z_E is approximately:

$$Z_E = \frac{1}{g_m} + \frac{R_B + r_b}{\beta} \quad (4.13)$$

If we consider an E60X device biased at $I_C = 10$ mA (as in Section 4.2), we have $g_m = 0.39$ A/V, $\beta = 80$ and $r_b = 57$ ohm. If $R_S = 0$, the minimum emitter impedance is 3.3 ohm. When Equation (4.12) is written in canonical form, the high-frequency behaviour of Z can be studied. The response consists of a zero and a pole:

$$Z_E = \frac{K(1 + s\tau_z)}{1 + s\tau_p} \quad (4.14)$$

$$K = \frac{G + g_{\pi}}{G(g_m + g_{\pi})} \quad (4.15)$$

$$\tau_z = \frac{C_{\pi}}{G + g_{\pi}} \quad (4.16)$$

$$\tau_p = \frac{C_{\pi}}{g_m + g_{\pi}} \quad (4.17)$$

Considering again the E60X device biased at $I_C = 10$ mA and $R_S = 0$, we find: $G = 1/r_b = 0.0175$ A/V, $g_m = I_C/v_t = 0.39$ A/V, $g_{\pi} = g_m/\beta = 0.0048$ A/V, $C_{\pi} = C_j + C_D = 0.4$ pF + 6.2 pF = 6.6 pF. With these values, we find $f_z = 1/(2\pi\tau_z) = 538$ MHz and $f_p = 1/(2\pi\tau_p) = 9520$ MHz. In other words, the emitter impedance is substantially inductive over much of the useful frequency range of the device. Since $f_p \gg f_z$, we can ignore the pole and model the impedance as a series RL circuit:

$$Z_E = R_E + sL_E \quad (4.18)$$

$$R_E = K = \frac{G + g_{\pi}}{G(g_m + g_{\pi})} \quad (4.19)$$

$$L_E = K\tau_z = \frac{C_{\pi}}{G(g_m + g_{\pi})} \quad (4.20)$$

The situation is slightly more complicated if the collector resistance is nonzero, and much more complicated if the full large-signal model of the transistor is considered. The basic behaviour, however, remains the same: The common base input impedance is substantially inductive, and both R and L are nonlinear through $g_m(I_C)$, $C_{\pi}(I_C)$ and $G(r_b(I_C))$. Any increase in the input impedance will cause an increase in the ac voltage across r_{π} and

c_π and a resulting increase in compression. In extreme cases, the contribution of the base-emitter junction to overall compression can be comparable to the collector-base contribution.

For an E60X biased at 10 mA emitter current and with a collector resistance of 200 ohms, the external base resistance R_S is varied from zero to 100 ohms (see Figure 4-17).

Figure 4-17 : Test circuit to determine the effect of R_S on transresistance linearity

The unloaded common base transresistance is simulated to find the compression and intermodulation performance with a 9 mA peak current input. Note that the E60X device has a 57 ohm parasitic bulk base resistance. As R_S is increased, gain magnitude compression increases, but phase compression and intermodulation distortion decrease:

Table 4.3: Distortion versus R_b for Common-Base Amplifier

RS	ΔM (mdB)	$\Delta \Phi$ (deg)	IM3 (dB)
0	27.5	1.250	63.2
50	30.15	1.235	63.4
100	32.65	1.220	63.6

Depending on bias current and collector resistance, the compression and intermodulation distortion may increase or decrease with base resistance.

The effect of base resistance on common base and cascode stages is interesting for three reasons. First, there is always parasitic bulk base resistance in any bipolar transistor, and especially in high-speed IC transistors. The effect of this resistance must be fully understood so that device size and bias conditions can be selected intelligently. Second, on-chip common base bias voltages are often provided by high-impedance resistive dividers from the supply voltage rails as shown in Figure 4-18a. This helps to reduce unwanted coupling between stages and minimizes power consumption, but introduces an effective base resistance of $R_1 R_2 / (R_1 + R_2)$. Third, series resistance is sometimes intentionally added to a differential cascode bias network as shown in Figure 4-18b to provide first-order compensation for β variation with collector current [53], or to increase common-mode stability in the transient response [54]. Note that although the sum of base currents I_{B1} and I_{B2} is ideally constant, the variation in β with collector current causes some variation in this sum, and therefore in the divider bias voltage. The presence of R_{B1} and R_{B2} reduces the effect of this variation.

Figure 4-18 : Resistor dividers for cascode bias

The effect of nonideal current sinks is of interest in level shifter stages and output driver stages. In both cases we are interested in the output impedance of the current sink,

and its behaviour as a function of amplitude and frequency.

In the simplest case, the sink will be a single transistor running at collector current I_C with some equivalent external base resistance R_B and emitter resistance R_E (see Figure 4-19a). In Figure 4-19b, the transistor has been replaced by its hybrid-pi equivalent circuit, and a test current is applied at the collector.

Figure 4-19 : Collector impedance of a bipolar transistor

When the circuit is solved using nodal analysis, the impedance looking into the collector is $Z_C = v_3/i_x$. Neglecting C_μ , this is found to be:

$$Z_C = \frac{G(g_m + g_\pi + sC_\pi) + (G + g_\pi + sC_\pi)(g_E + g_o)}{g_o((G + g_E)(g_\pi + sC_\pi) + g_E G)} \quad (4.21)$$

where $g_E = 1/R_E$, $g_o = 1/r_o$ and all other variables are defined as for Equation (4.12). For low frequencies and $\beta \gg 1$, we find Z_C is approximately:

$$Z_C = r_o(1 + g_m R_E) \quad (4.22)$$

The small-signal output resistance r_o is a first-order approximation to account for the Early effect (basewidth modulation due to collector-base voltage). Looking again at the E60X device running at 10 mA, r_o is calculated as $r_o = V_A/I_C = 1870$ ohms. For an E60X

device biased at 10 mA with $R_E = 50 \text{ ohm}$, the impedance seen at the collector is over twenty times higher than for $R_E = 0$. Note that Equation (4.22) implies that Z_E is independent of G (the base circuit impedance). This is not strictly true, but is a useful low-frequency approximation.

To analyze the behaviour at high frequency, Equation (4.21) can be expressed in canonical form:

$$Z_C = \frac{K(1 + s\tau_z)}{1 + s\tau_p} \quad (4.23)$$

$$K = \frac{G(g_\pi + g_m) + (G + g_\pi)(g_E + g_o)}{g_o(g_\pi(g_E + G) + g_E G)} \quad (4.24)$$

$$\tau_z = \frac{C_\pi(G + g_E + g_o)}{G(g_\pi + g_m) + (G + g_\pi)(g_E + g_o)} \quad (4.25)$$

$$\tau_p = \frac{C_\pi(g_E + G)}{g_\pi(g_E + G) + g_E G} \quad (4.26)$$

Using the same values as before for G , g_m , g_π and C_π , and setting $g_E = 0.02 \text{ A/V}$ and $g_o = 0.0005 \text{ A/V}$, τ_z and τ_p can be evaluated. The corresponding pole and zero frequencies are $f_z = 4670 \text{ MHz}$ and $f_p = 341 \text{ MHz}$. The current source output impedance is therefore capacitive over the useful range of the device. Since $f_z \gg f_p$, the zero can be neglected and the output admittance Y_C can be expressed:

$$Y_C = G_C + sC_C \quad (4.27)$$

$$G_C = \frac{1}{K} = \frac{g_o(g_\pi(g_E + G) + g_E G)}{G(g_\pi + g_m) + (G + g_\pi)(g_E + g_o)} \quad (4.28)$$

$$C_C = \frac{\tau_p}{K} = \frac{C_{\pi} g_o (g_E + G)}{G (g_{\pi} + g_m) + (G + g_{\pi}) (g_E + g_o)} \quad (4.29)$$

Both G_C and C_C vary with amplitude and frequency through g_m , g_{π} and C_{π} . The exact behaviour of the current sink using large-signal models is more complex, but similar to what is expected using the above analysis. The circuit of Figure 4-20 was used to examine the compression and intermodulation performance of ideal and nonideal current sinks. The unloaded common collector stage is tested first with an ideal 10 mA current sink (Figure 4-20a), then with a more realistic current mirror (Figure 4-20b).

Figure 4-20 : Test circuit to determine the effect of real and ideal current sinks on distortion

The signal is 2 V peak at 140 MHz for the compression test, and 1.41 V peak each at 130 MHz and 150 MHz for the intermodulation test. Although gain magnitude compression is lower for the real sink, the phase compression and intermodulation distortion

are higher:

Table 4.4: Distortion for Ideal and Real Current Sinks

sink	ΔM (μdB)	$\Delta\Phi$ (deg)	IM3 (dB)
ideal	672	0.016	93.4
real	165	0.063	82.6

Careful control of the emitter impedance of the sink transistor allows adjustment of the magnitude and phase compression characteristics.

Superior current sinks can be designed using the Wilson topology (Figure 4-21a) or cascode topology (Figure 4-21b). These circuits provide much higher output impedance than sinks with simple resistor degeneration. With both G_C and C_C reduced (Equations (4.28) and (4.29)), the output impedance has less dependence on both amplitude and frequency. These more sophisticated circuits demand at least one more base-emitter voltage drop in headroom, which may be unacceptable in a low-voltage application.

Figure 4-21 : Wilson (a) and cascode (b) current sinks

4.6 Device Mismatches

In Section 4.4, the advantages of differential circuit topologies were explained. The reduction of odd-order intermodulation and the elimination of even-order intermodulation are very desirable, but the designer must always consider the degradations caused by device mismatches.

Modern bipolar processes feature superb control over etching, diffusion, implantation and annealing. As a result, both vertical and lateral features are kept very uniform. The greatest variations in device parameters arise from the lithography and processing variations associated with the bases and especially the emitters of the NPN transistors. Because these are the smallest lateral features and the shallowest vertical features, a variation of (say) 100 angstroms in an emitter dimension will have a much more noticeable effect than the same variation in a collector buried layer dimension.

The first step in modelling the effect of device mismatches is to translate variations in process parameters to variations in device electrical parameters. Most authors assume that normally distributed lithography variations produce normal distributions of device parameters. If the mean and variance of each parameter value are known, the statistics for all circuit variables can be calculated.

A simple example will illustrate the procedure. The circuit of Figure 4-22 is a differential transconductance, and we wish to know how the transconductance is affected by normally distributed variations in device parameters.

Figure 4-22 : Test circuit to determine the effect of mismatches on $(I_1 - I_2)$

The governing equations for the circuit are:

$$(V_{i1} - V_{i2}) - (V_{BE1} - V_{BE2}) = \frac{I_1}{\alpha_1} R_1 - \frac{I_2}{\alpha_2} R_2 = \Delta V \quad (4.30)$$

$$\frac{I_1}{\alpha_1} + \frac{I_2}{\alpha_2} \approx I_T \quad (4.31)$$

From these we can obtain $I_1 - I_2$ directly:

$$I_1 - I_2 = \frac{(\alpha_1 - \alpha_2) \Delta V + I_T (\alpha_1 R_2 - \alpha_2 R_1)}{R_1 + R_2} \quad (4.32)$$

The mean and variance of $I_1 - I_2$ can be calculated using knowledge of the means and variances of α , R , I_T and ΔV (which includes base-emitter voltage mismatch). The treatment of functions of independent random variables is straightforward [55] and is summarized in Appendix F. In reality, many of the SPICE parameters are not independent; for example, forward current gain βF and bulk base resistance R_B are strongly correlated. For this reason, computer modelling is used extensively to predict worst-case offset effects. The most common (but time consuming) method of statistical design is

Monte Carlo simulation, where many combinations of allowable offsets are incorporated into the simulation run, and the resulting distribution of some performance measure is observed.

An alternative method which uses less computing resources is to determine the combination of offsets and mismatches which most strongly affects performance. Some finesse is needed in choosing the tolerable amount of offset for each parameter, because choosing a succession of "three-sigma" offsets for parameters will almost certainly indicate that any design is impossible to produce with acceptable yields. After trying a variety of methods, it was noticed that most offset calculations are dominated by a few terms, so most of the design effort was concentrated on these elements.

Insufficient attention to offset control will cause an increase in phase compression and intermodulation. Matching problems are aggravated by temperature drift of parameters, especially by increasing temperature. The equation relating V_{BE} mismatch for two devices is:

$$\Delta V_{BE} = V_{BE1} - V_{BE2} = \frac{kT}{q} \ln \left(\frac{J_{C1}}{J_{C2}} \right) \quad (4.33)$$

where J_C is collector current density. If equal current flows in each device and their emitter areas are mismatched then:

$$\Delta V_{BE} = \frac{kT}{q} \ln \left(\frac{A_{E2}}{A_{E1}} \right) \quad (4.34)$$

where A_E is emitter area. Note that the offset voltage depends weakly on the area mismatch, but is proportional to temperature. The "three-sigma" matching for the devices used in this work is given in the following table. The data are for matching over a 10 cm wafer at 25 degrees C.

Table 4.5: Matching Data for SH3 Devices

Parameter	3σ
E15 NPN (ΔV_{BE})	2.0 mV
E60 NPN (ΔV_{BE})	1.8 mV
E120 NPN (ΔV_{BE})	1.0 mV
Lateral PNP (ΔV_{BE})	2.0 mV
NPN β ($\Delta\beta/\beta$)	5 %
PNP β ($\Delta\beta/\beta$)	8 %
P+ Resistor (matching)	1 %
AB Resistor (matching)	1 %
Nichrome Resistor (matching)	1 %

As expected, physically larger transistors have better matching because some absolute lithography error will have less relative effect for a larger transistor. Although the P+ and AB resistors are made from different implants, they are about the same size and shape in lateral dimensions, so they have similar matching tolerances. For best matching, all emitter stripes on a chip should run in the same direction and all resistors of the same type should run in the same direction. Devices which are intended to be matched pairs are chosen to be as close together physically as possible. This leads to matching much tighter than the values implied in the table above. In fact, emitter area matching for adjacent devices is typically about 1 %.

The table below is intended to show the effect of emitter mismatch on distortion performance. The circuit of Figure 4-23 is simulated with attention to four transconductance distortion measures: magnitude and phase compression and second-order and third-order intermodulation. The simulations are performed for emitter area mismatches of from zero to five percent.

Figure 4-23 : Test circuit to determine the effect of mismatches on g_m nonlinearity

The input for the compression test is 200 mV peak at 140 MHz, while the inputs for the intermodulation test are 100 mV peak each at 130 and 150 MHz.

Table 4.6: Distortion as a Function of Emitter Area Mismatch

A_{E1}/A_{E2}	ΔM (mdB)	$\Delta\Phi$ (deg)	IM2 (dB)	IM3 (dB)
1.00	700.39	0.30355	-	48
1.01	698.86	0.30501	92	48
1.02	697.34	0.30647	86	48
1.03	695.84	0.30792	83	48
1.04	694.36	0.30937	80	48
1.05	692.89	0.31081	78	48

Amplitude compression changes by about 0.01 dB and phase compression changes by about 0.01 degree. Third-order intermodulation is fairly high, and is therefore insensitive to mismatch, but the second-order intermodulation is strongly affected by mismatches of even a few percent. While the actual distortion contributions for this example are very small, it must be remembered that these are relatively large devices biased at relatively high current. Mismatch and distortion will be greater for small devices biased at low

currents. The distortion also adds up for each differential pair in the signal path, and the number of such pairs may be ten or more in a signal processing block. In general, individual transistor parasitics most strongly affect odd-order intermodulation but differential circuit matching affects even-order intermodulation.

4.7 Thermal Distortion and Temperature Effects

SPICE is fairly accurate for most simulations of the effect of temperature on bipolar circuit dc behaviour. For example, SPICE will accurately predict the dc performance of bias circuits such as bandgap voltage references and current mirrors over a wide range of ambient temperatures.

All SPICE analyses have one feature in common: the device temperature is used to calculate the parameter values only during the initial dc solution. After this, temperature has no further effect on the rest of the analysis. This is entirely suitable for a bias analysis (where nothing happens to the circuit after the initial dc solution), or for small-signal ac and noise analyses (where signal swings are very small and device operating points are not greatly perturbed). A problem arises for large-signal transient and quasistatic compression analyses, however: large signal swings on a transistor cause dynamic changes in the device power dissipation, and therefore in its temperature. A change in temperature can mean large parameter value changes, which in turn affects the device electrical performance. In effect, a "thermal feedback" mechanism acts on the device, and this feedback is not accounted for in the SPICE GP2 model. In addition to "self heating", a device may also experience "cross heating" from neighboring devices with varying power dissipation.

Electrothermal interaction has been studied for some time [56,57,58], and a variety of methods have been used to account for it. In principle, the transistor could be simulated using a four-dimensional finite element method (FEM) or another numerical technique [59]. After describing the geometry, doping and thermal properties for each element, the solver could handle the carrier diffusion, drift and continuity equations as well as the heat

flow equation to predict the response to electrical transients applied at the terminals. This approach is very rigorous, but far too slow and cumbersome for circuit design except for the very smallest circuits. It also suffers from convergence problems. It is also possible to apply a frequency-domain solution to correct small-signal Y parameters [60]. This method provides good small-signal accuracy and is well-suited to CAD implementation because of its simplicity.

The problem of dynamic thermal distortion is more complicated than the small-signal case. FEM or a similar method can be used, but speed and convergence remain difficult problems. For this work, a simple method was chosen to simulate coupling of the thermal and electrical problems without excessive loss of speed or robustness. The basic strategy is to use the electrical-thermal analogy so that SPICE can directly solve the thermal problem simultaneously with the electrical one. In this analogy, temperature is analogous to voltage, power flow is analogous to current and thermal resistance is analogous to electrical resistance. Each device is modelled as shown in Figure 4-24.

Figure 4-24 : Thermal model for NPN bipolar transistor

The governing equations for the model are:

$$V_1 = V_{CE}I_C - V_0I_0 \quad (4.35)$$

$$V_2 = V_1 \exp\left(\frac{-t}{\tau_{th}}\right) \quad (4.36)$$

$$V_{BEX} = V_{BE} + kV_2 \quad (4.37)$$

$$k = \left(\frac{\partial V_{BE}}{\partial T}\right)\left(\frac{\partial T}{\partial P}\right) \quad (4.38)$$

where (V_{CE}, I_C) is the instantaneous operating point of interest, (V_0, I_0) is the bias point where the models were derived or measured, τ_{th} is the thermal time constant of the device, $\partial V_{BE}/\partial T$ is about -1.6 mV/degree C for these devices and $\partial T/\partial P = R_{th}$ is the thermal resistance of the device. For the three transistors used in this work, the relevant data are:

Table 4.7: Thermal Properties of SH3 Devices

Device	R_{th} (deg C/W)	τ_{th} (ns)
E15	200	50
E60X	87	100
E120X	60	200

It is apparent that the thermal feedback loop will have a low-pass characteristic with a single pole no higher than 20 MHz. Obviously this kind of behaviour is very important for baseband amplifiers such as ADC drivers, but it is also relevant for IF amplifiers because the rolloff with frequency is slow. For the QAM system discussed in Chapter 2, the IF is centred at 140 MHz so if the pole is at 20 MHz, the magnitude of the response is:

$$|H(f)| = \left| \frac{H_0}{1 + j\left(\frac{f}{f_p}\right)} \right| = H_0 \left| \frac{1}{1 + j\left(\frac{140}{20}\right)} \right| = 0.14H_0 \quad (4.39)$$

In other words, the thermal distortion at IF is only 17 dB down from its maximum baseband contribution.

One problem with this modelling technique is that the thermal time constants are three to four orders of magnitude larger than the device electrical time constants. This makes the system of circuit equations mathematically "stiff", causing excessive simulation run times or a failure to converge. It is possible to examine circuit behaviour qualitatively by using thermal time constants much shorter than the real values but much longer than the device electrical time constants (by one or two orders of magnitude).

In order to determine the effects of thermal distortion, the simple circuit of Figure 4-25 is simulated with and without thermal feedback as shown in Figure 4-24.

Figure 4-25 : Test circuit to determine the effect of thermal distortion on compression and intermodulation

Magnitude and phase compression at 140 MHz are predicted, as well as second-order and third-order intermodulation levels with 130 and 150 MHz test tones. The results are as follows:

Table 4.8: Effect of Thermal Modelling on Simulated Distortion

Case	CP(1 dB)	CP(0.1 dB)	IM2	IM3
Standard model	-5.61 dBm	-4.9 dBm	30.0 dB	40.3 dB
Thermal model	-5.69 dBm	-4.56 dBm	29.9 dB	40.0 dB

Even for this low-power class-A amplifier, thermal distortion has a measurable effect. Usually it is advisable to simulate thermal distortion for the output devices in a power amplifier, or for any device whose instantaneous power dissipation varies by more than 50% of its quiescent dissipation.

4.8 Feedback

Most introductory circuit design texts emphasize the beneficial effects of feedback for improving gain stability, bandwidth and distortion performance in amplifiers. In Figure 4-26a, the simplest feedback topology is shown with input $X(f)$ and output $Y(f)$, forward transfer function $A(f)$ and feedback transfer function $B(f)$.

Figure 4-26 : Negative feedback topology (a) and its effect on distortion (b)

The transfer function is:

$$H(f) = \frac{Y(f)}{X(f)} = \frac{A(f)}{1 + A(f)B(f)} \quad (4.40)$$

If the loop gain $A(f)B(f) \gg 1$, the transfer function will be approximately $H(f) = 1/B(f)$. Usually $|A(f)|$ is large but not well-controlled; for example it may represent the gain of an amplifier. Normally $B(f)$ represents the transfer function of a simple passive network, for example a resistive divider. Since $B(f)$ is precisely controlled, nonideal effects from $A(f)$ will be reduced. Suppose the amplifier has a nonlinear output stage which adds distortion d as shown in Figure 4-26b. The output Y is then:

$$Y = \frac{AX}{1 + AB} + \frac{d}{1 + AB} \quad (4.41)$$

In other words, the distortion degradation outside the loop is the distortion inside the loop divided (roughly) by the loop gain.

Feedback around a single transistor is implemented using two techniques: series feedback and shunt feedback. These configurations are shown in Figures 4-27a and 4-27b, respectively.

Figure 4-27 : Single-transistor amplifiers with series (a) and shunt (b) feedback

Using the simplest possible hybrid-pi model, the voltage gain expressions for these two circuits are:

$$\left(\frac{V_o}{V_i} \right)_{series} = \frac{-g_m R_C}{1 + \frac{g_m}{g_\pi + g_E + sC_\pi}} \quad (4.42)$$

$$\left(\frac{V_o}{V_i} \right)_{shunt} = \frac{\frac{R_C}{R_F} - g_m R_C}{\frac{R_C}{R_F} + 1} \quad (4.43)$$

In each case, the gain without feedback would be $-g_m R_C$. If both series and shunt feedback are used simultaneously as in Figure 4-28, the gain expression becomes:

$$\left(\frac{V_o}{V_i} \right)_{series-shunt} = \frac{-g_m (R_F || R_C)}{1 + g_m R_E} \quad (4.44)$$

where $x || y = xy/(x+y)$.

Figure 4-28 : Single-transistor amplifier with combined series-shunt feedback

All three of these classical configurations increase bandwidth, stabilize gain against temperature variations and reduce distortion due to transconductance nonlinearity. Unfortunately, they all increase noise figure over that of the basic common emitter stage, but they are simple to apply in situations other than low-noise stages.

The single transistor stages illustrate the concept of "local" feedback; the feedback

path is around a single device and corrects fairly crudely for the nonideal properties of that device. Greater distortion improvement and smaller degradation of noise figure can be attained [61,62,63] by employing feedback around two or more devices as shown in Figure 4-29.

Figure 4-29 : Multiple-transistor feedback amplifiers

This illustrates the concept of "global" feedback which is used to great advantage in operational amplifiers. The problem with global feedback is that the gain and phase margins are reduced by the large loop gain magnitude and accumulating phase shift which occurs for even a modest number of high-quality devices. For operational amplifiers, the problem is solved by compensating with one or two dominant poles at low frequency to guarantee less than 0 dB loop gain at 180 degrees of phase shift. This is

acceptable for an amplifier whose application will be at or near dc, but will not work at all for a true wideband amplifier. Shaping networks are sometimes used to provide phase margin relief, but the component value sensitivities are too great to allow a reproducible IC design.

One technique which dramatically improves linearity with only minimal penalties in noise, bandwidth and power consumption is the use of "three-transistor buffers" (3TBs).

Figure 4-30 : Single transistor and corresponding three-transistor buffer (3TB)

In Figure 4-30a, a typical transistor is shown with current sink bias I_{1a} . The collector, base and emitter are labelled C, B and E respectively. In Figure 4-30b, an equivalent 3TB topology is shown with corresponding terminals C', B' and E'. Q1 is now embedded in a high-gain feedback loop completed by differential pair Q2-Q3 and current mirror M. The differential pair is biased by current sink I_{2b} and functions as a common collector/common base pair, with open-loop gain being determined by the transconductance of Q3 and the load impedance seen at its collector. The closed-loop gain is unity. Typically Q2 and Q3 will be smaller devices than Q1, and will be biased with proportionately lower current.

The 3TB arrangement has several advantages over the simpler one. First, since Q2 is smaller than Q1 and runs at lower current, the input impedance of the 3TB is higher than that of the single transistor. The input capacitance in particular is reduced, and any Miller effect at node B is eliminated at B'. Second, reverse isolation from C' to B' is much greater than from C to B; this improves the stability margin of amplifiers and reduces unwanted coupling. Third, all nonlinearity contributions inside the 3TB loop are divided by the loop gain when they appear at the output. This includes all electrical and thermal contributions from the base-emitter circuit of Q1, which often dominate the distortion of useful circuits. Of course, extra devices must be used and current sink I_{2b} increases power dissipation, but the linearity of the 3TB with bias sinks I_{1b} and I_{2b} will be superior to that of the single device biased at $I_{1a} = I_{1b} + I_{2b}$. The designer can therefore provide better linearity at the same dissipation, or the same linearity at lower dissipation. If Q1 in the 3TB can be run at lower collector current than the single device, its shot noise contribution to overall noise will be lower. A typical implementation of the 3TB is shown in Figure 4-31b (as well as the equivalent single device in Figure 4-31a).

Figure 4-31 : Test circuits to compare magnitude and phase compression

The two circuits were evaluated for gain compression at 140 MHz. The results are shown in Figure 4-32. Note that phase compression for the 3TB is worse, but magnitude compression is better than for the basic configuration. The magnitude and phase compression can be adjusted by varying the tail currents in the 3TB.

Figure 4-32 : Magnitude and phase compression for the circuits of Figure 4-31

One feedback topology which is being used frequently to reduce distortion is "current feedback". A comparison of current feedback with the more conventional voltage feedback illustrates the difference between them.

Figure 4-33 : Voltage feedback amplifier topology

In Figure 4-33, a voltage feedback amplifier is shown in a noninverting feedback configuration. Both amplifier inputs are high impedance nodes, and the device has a large voltage gain:

$$A_v(s) = \frac{V_{out}}{V_{in} - V_x} \quad (4.45)$$

The absence of amplifier input current at node x allows a simple calculation of ideal closed-loop gain:

$$\frac{V_{out} - V_{in}}{R_F} = \frac{V_{in}}{R_G} \quad (4.46)$$

$$\frac{V_{out}}{V_{in}} = G = 1 + \frac{R_F}{R_G} \quad (4.47)$$

Now assume that $A_v(s)$ can be expressed as the ratio of zero and pole polynomials:

$$A_v(s) = \frac{Z_v(s)}{P_v(s)} \quad (4.48)$$

Then a new expression for V_{out}/V_{in} is derived in terms of the closed-loop gain and $Z_v(s)$ and $P_v(s)$:

$$V_{out} = A_v (V_{in} - V_x) = A_v \left(V_{in} - \frac{V_{out} R_G}{R_F + R_G} \right) \quad (4.49)$$

$$\frac{V_{out}}{V_{in}} = \frac{G Z_v(s)}{Z_v(s) + G P_v(s)} \quad (4.50)$$

Note that the pole locations of the transfer function are determined by the gain G . The effect of this is the well-known fixed gain-bandwidth product, which is a fundamental limitation of this type of circuit. The linearity implications of Equation 4.50 also arise from the pole dependency on G .

Figure 4-34 : Current feedback amplifier topology

In contrast, the current feedback amplifier shown in Figure 4-34 has one high im-

pedance input node (in) and one low impedance input node (x). In this case the device transresistance is large:

$$A_r(s) = \frac{V_{out}}{I_x} \quad (4.51)$$

Since the unity-gain buffer forces V_{in} to appear at node x, the closed-loop gain is also described by Equation 4.47. If $A_r(s)$ is described as the ratio of its zero and pole polynomials:

$$A_r(s) = \frac{Z_r(s)}{P_r(s)} \quad (4.52)$$

then this transfer function can also be expressed in terms of the closed-loop gain and $Z_r(s)$ and $P_r(s)$:

$$V_{out} = A_r I_x = A_r \left(\frac{V_{in}}{R_G} - \frac{(V_{out} - V_{in})}{R_F} \right) \quad (4.53)$$

$$\frac{V_{out}}{V_{in}} = \frac{G Z_r(s)}{Z_r(s) + R_F P_r(s)} \quad (4.54)$$

In this case, the transfer function poles are dependent on R_F rather than G . In other words, if R_F is held constant and R_G is reduced, gain will be increased and bandwidth will remain constant. Since bandwidth remains constant regardless of gain, distortion performance is improved over the voltage feedback case. Although real amplifiers cannot achieve the ideal performance described above, many commercial parts use this principle to great advantage.

4.9 Feedforward

Besides feedback, another useful strategy is feedforward. The principle, as illustrated in Figure 4-35, is to sample the main amplifier input and output signals, scale and delay them appropriately using networks B_1 and B_2 , and compare them. The resultant error signal is scaled and delayed by network B_3 and added to the main amplifier output in such a way as to cancel the degradations produced by the main amplifier.

Figure 4-35 : Feedforward amplifier topology

The transfer function is:

$$H(f) = \frac{Y(f)}{X(f)} = A(f) (1 + B_2(f) B_3(f)) - B_1(f) B_3(f) \quad (4.55)$$

Feedforward is a technique well-suited to IC design. The importance of matched gains and delays in the various branches is paramount, and so is good thermal tracking. One of the most common forms of wideband feedforward compensation is the "cascomp" (cascode compensator) [64,65]. The cascomp topology is shown in Figure 4-36.

degraded phase compression performance. The cascomp amplifier is also noisier than the cascode because of the low series resistance in the "loop of base-emitter drops" (Q3-Q6) in Figure 4-37b. This issue is addressed further in Section 4.10.

Figure 4-37 : Cascode and cascomp test circuits to be compared for gain magnitude and phase compression

Figure 4-38 : Magnitude and phase compression for the circuits of Figure 4-37

4.10 Translinear and Current-Mode Circuits

There are three fundamental operations in linear analog signal processing: scaling by a fixed constant (amplification or attenuation), scaling by a variable (two- or four-quadrant multiplication) and summing. For each of these operations, there is a variety of suitable implementations for monolithic bipolar ICs. The choice of implementation is determined by noise, linearity, bandwidth and power dissipation requirements. The most effective implementation for wideband high-linearity analog signal processing is the current-mode strategy.

In a typical current-mode circuit, input signal voltages are converted to currents

using some type of transconductance, then all required processing is performed on the currents. Finally, the output current is converted back to a voltage using a transresistance, and optionally buffered off-chip using some kind of voltage follower. For example, in Figure 4-39 a typical processing function is required; $y = a_1 x_1 + kx_2$, where x_1 and x_2 are signal inputs, a_1 is a variable coefficient and k is a constant.

Figure 4-39 : Functional implementation of $y = a_1 x_1 + Kx_2$

One possible circuit implementation is shown in Figure 4-40. Q1, Q2, R1 and R2 make up the x_1 input transconductor, while Q3, Q4, R3 and R4 are for the x_2 input. Q5-Q8 are connected as a four-quadrant multiplier, with coefficient a_1 applied as a differential voltage at the bases. Q9-Q12 are connected as a current amplifier with gain k . The output currents from the multiplier and amplifier are summed at a pair of low impedance nodes (the emitters of common base amplifier Q13-Q14). The currents are converted back to voltages here (using R5-R6), and buffered off through emitter followers Q15 and Q16.

Figure 4-40 : Circuit implementation of $y = a_1x_1 + Kx_2$

The current-mode approach has several key advantages [66]. First, wide bandwidth is maintained because all devices are operated effectively as common base or common collector amplifiers. These have inherently wider bandwidth than the common emitter configuration because the Miller effect is avoided. The frequency response is usually totally determined by a single dominant pole due to the collector resistance and collector capacitance of the common base amplifier (Q13-Q14 in the example). A second and related advantage is superior linearity. Because large collector-base voltage swings are avoided everywhere except at the common base output, the collector-base junctions make no significant contributions to nonlinearity. Also, these circuits take advantage of the fact that the current gains $\alpha(I_C)$ and $\beta(I_C)$ are much weaker nonlinearities than the transcon-

ductance $g_m(V_{BE})$. The stacked topology as shown in Figure 4-41a is very efficient in the sense that dc bias currents are "reused" by all blocks, but the required voltage headroom for the power supplies can be excessive if too many levels are stacked. Lower voltage supplies can be used if a folded topology is selected as shown in Figure 4-41b. Note that high-quality PNP transistors are necessary for efficient implementation of folded current-mode circuits.

Figure 4-41 : Conventional (a) and folded (b) cascode amplifiers

A large and useful class of current-mode circuits is the set of translinear circuits; that is, circuits which make use of the linear dependence of transconductance on collector current. This linear relationship holds accurately for six or more decades of collector current variation, and is therefore a good basis for ultralinear circuit design. Physically, true translinear circuits feature a closed loop of base-emitter voltage drops as shown in Figure 4-42.

Figure 4-42 : Translinear squaring circuit

If current I_0 is fixed, the output current I_y is proportional to the square of the input current I_x ; if I_x is fixed, I_y is proportional to the reciprocal of input current I_0 . The analysis proceeds from a knowledge of the base-emitter voltage drops in the circuit:

$$V_{BE1} + V_{BE2} = V_{BE3} + V_{BE4} \quad (4.57)$$

$$\frac{kT_1}{q} \ln \left(\frac{I_{C1}}{I_{S1}} \right) + \frac{kT_2}{q} \ln \left(\frac{I_{C2}}{I_{S2}} \right) = \frac{kT_3}{q} \ln \left(\frac{I_{C3}}{I_{S3}} \right) + \frac{kT_4}{q} \ln \left(\frac{I_{C4}}{I_{S4}} \right) \quad (4.58)$$

If the devices have identical geometry and there is no thermal distortion, then:

$$\ln \left(\frac{I_{C1}}{I_S} \right) + \ln \left(\frac{I_{C2}}{I_S} \right) = \ln \left(\frac{I_{C3}}{I_S} \right) + \ln \left(\frac{I_{C4}}{I_S} \right) \quad (4.59)$$

$$\ln\left(\frac{I_{C1}I_{C2}}{I_S^2}\right) = \ln\left(\frac{I_{C3}I_{C4}}{I_S^2}\right) \quad (4.60)$$

If $\beta \gg 1$ then $I_{C1} = I_{C2} = I_x$, $I_{C3} = I_0$ and $I_{C4} = I_y$:

$$I_x \cdot I_x = I_0 \cdot I_y \quad (4.61)$$

$$I_y = \frac{I_x^2}{I_0} \quad (4.62)$$

By using loops of transistors with different emitter area ratios, translinear circuits can be used to perform squaring, square root, geometric mean calculation, vector addition and subtraction, scaling and multiplication. The bandwidths of these circuits are very large, almost up to f_T for the devices used, and their operation remains linear over a huge range of signal levels. The drawbacks of translinear circuits are firstly, a strong sensitivity to thermal distortion and secondly, a tendency to be noisy.

The first difficulty is caused by the nature of the equation for base-emitter voltage drop as a function of collector current:

$$V_{BE} = \frac{kT}{q} \ln\left(\frac{I_C}{I_S}\right) \quad (4.63)$$

While errors in base-emitter geometry affect I_S directly, their effect on V_{BE} is through the relatively weak logarithmic dependence. On the other hand, V_{BE} is directly proportional to absolute temperature, and there is no emitter degeneration in the translinear loop, so device-to-device temperature differences will degrade circuit operation. For example, if $I_C = 1$ mA, $T = 300$ K and $I_S = 10^{-16}$ A, a 1% error in I_S causes a 257 μ V error in V_{BE} while a 1 degree temperature error causes a 2.58 mV error in V_{BE} . Fortunately, there

are several layout techniques which can minimize thermal distortion; these are discussed in Section 4.12.

The noise generation in translinear circuits arises from thermal and shot noise contributions in each device. Thermal noise is caused by random motion of carriers in regions of finite resistivity and shot noise is caused by statistical fluctuations in direct current crossing a potential barrier. Figure 4-43 shows the hybrid- π small-signal model of a bipolar transistor together with its conventional noise generators.

Figure 4-43 : Hybrid- π small-signal model of a bipolar transistor including noise generators

The three noise generators have values [67]:

$$\overline{v_b^2} = 4kTr_b\Delta f \quad (4.64)$$

$$\overline{i_c^2} = 2qI_C\Delta f \quad (4.65)$$

$$\overline{i_b^2} = 2qI_B\Delta f \quad (4.66)$$

When transistors are connected in undegenerated loops, the v_b^2 sources are connected in series, so the thermal contributions from the transistor bulk base resistances are large. The only real countermeasure against this noise degradation is to use large devices

(which have lower r_b than smaller devices). Larger devices have smaller parasitic resistances but larger capacitances. The bandwidth degradation in a properly-designed circuit will be minimal, however, because in current-mode operation there is very little ac voltage across these capacitances (except at the common base stage collectors).

In general, the use of translinear circuits provides an improvement in compression point but degrades noise figure. If the technique is to be useful, there must be a net improvement in dynamic range over conventional techniques. For comparison, the transconductances of Figure 4-44 are simulated for noise and compression point at 140 MHz.

Figure 4-44 : Transconductance noise/distortion test circuits

Figures 4-44a and 4-44b show a standard differential pair while Figure 4-44c shows a translinear approach. The results are as follows:

Table 4.9: Dynamic Range Comparison for Transconductances of Figure 4-44

Parameter	4-44a	4-44b	4-44c
CP(1 dB) _{in} (mV)	31	21	44
CP(1 deg) _{in} (mV)	29	26	61
Input noise (nV/rt Hz)	1.49	0.73	0.84

Table 4.9: Dynamic Range Comparison for Transconductances of Figure 4-44

Parameter	4-44a	4-44b	4-44c
Dynamic range Δ (dB)	0	2.8	8.0

Since there is an improvement in dynamic range, the translinear technique is useful in this case. In general, translinear circuit design is governed by a careful choice of device sizes and bias currents. This is especially true for multipliers, which are often the dominant noise contributors in analog signal processing circuits.

4.11 Distortion and Noise in Multipliers

In the context of this work, a multiplier is defined to have one signal input port, one signal output port and a gain control input whose bandwidth is usually substantially less than that of the signal ports. Multipliers are commonly used in automatic gain control (AGC) circuits, transversal equalizers and clock recovery circuits.

Figure 4-45 : Classical wideband multiplier

The classic four-quadrant multiplier topology is shown in Figure 4-45. Signal input

v_y is multiplied by control signal v_x to produce output signal v_z :

$$V_z = K V_x V_y \quad (4.67)$$

The scaling constant K (with units $1/V$) is easily determined from a perturbation analysis. First, assume infinite β and consider the X and Y inputs:

$$\Delta V_{BE_x} = V_{BE2} - V_{BE1} = \frac{kT}{q} \ln \left(\frac{I_{C2}}{I_{C1}} \right) = \frac{kT}{q} \ln \left(\frac{I + \Delta I_x}{I - \Delta I_x} \right) \quad (4.68)$$

$$\Delta V_{BE_y} = V_{BE3} - V_{BE4} = \frac{kT}{q} \ln \left(\frac{I_{C3}}{I_{C4}} \right) = \frac{kT}{q} \ln \left(\frac{I + \Delta I_y}{I - \Delta I_y} \right) \quad (4.69)$$

We note that ΔV_{BE_x} will also be present for the pair Q5-Q6, and hence for Q7-Q10 and Q8-Q9:

$$\Delta V_{BE_x} = V_{BE2} - V_{BE1} = V_{BE6} - V_{BE5} = V_{BE7} - V_{BE10} = V_{BE8} - V_{BE9} \quad (4.70)$$

Transistors Q7-Q10 are the multiplier quad. Q7 and Q10 divide I_{C3} according to ΔV_{BE_x} :

$$\frac{I_{C7}}{I_{C10}} = \frac{I_{C2}}{I_{C1}} = \exp \left(\frac{q \Delta V_{BE_x}}{kT} \right) = \frac{I + \Delta I_x}{I - \Delta I_x} \quad (4.71)$$

and since $I_{C3} = I_{C7} + I_{C10} = I + \Delta I_y$, we can solve for I_{C7} and I_{C10} :

$$I_{C7} = \frac{(I + \Delta I_y)(I + \Delta I_x)}{2I} \quad (4.72)$$

$$I_{C10} = \frac{(I + \Delta I_y)(I - \Delta I_x)}{2I} \quad (4.73)$$

Similarly, for the inner pair:

$$I_{C8} = \frac{(I - \Delta I_y)(I + \Delta I_x)}{2I} \quad (4.74)$$

$$I_{C9} = \frac{(I - \Delta I_y)(I - \Delta I_x)}{2I} \quad (4.75)$$

Now we note that currents are summed at the collectors and differential output voltage V_z appears across the collector loads:

$$V_z = R_z((I_{C7} + I_{C9}) - (I_{C8} + I_{C10})) = \frac{2R_z}{IR_xR_y} V_x V_y \quad (4.76)$$

Therefore the multiplier gain constant $K = 2R_z/(IR_xR_y)$.

As a translinear circuit with current inputs and outputs, the multiplier quad is highly linear over decades of current variation. The main deviations from ideality are due to variation of β with collector current, parasitic base resistance and emitter area mismatch [68]. The quads are also susceptible to thermal distortion [69]; in fact this is frequently the dominant source of nonlinearity for a multiplier. All these considerations can be minimized by the use of large devices and careful layout techniques.

A useful inherent property of this type of multiplier is that the devices which carry the largest signal current also run at the largest bias current. In Figure 4-46a, steering voltage $V_A(+) = V_A(-)$; this corresponds to a multiplier weight of zero, or a nulled condition. Each device operates at an equal emitter current of $I/2$.

In Figure 4-46b, $V_A(+)-V_A(-) > 5kT/q$ so the right-hand devices in the quad are cut off and all the signal current is carried by the left-hand devices. These effectively form a common base pair biased at emitter currents I . Figure 4-46c shows the circuit when $V_A(-)$

$V_A(+)>5kT/q$; in this case the left-hand devices are cut off, the emitters are "crossed" and the output will have the same magnitude and the opposite sign to that in Figure 4-46b.

Figure 4-46 : Multiplier quad with bias currents at null (a), maximum positive gain (b) and maximum negative gain (c)

In Figure 4-47 the output differential current is $\Delta I_y = I_{o1} - I_{o2}$ and the input differential current is $\Delta I_x = I_{i1} - I_{i2}$.

Figure 4-47 : Multiplier test circuit

If the quad is unbalanced by applying differential voltages $\pm V_A/2$, then ΔI_y versus ΔI_x can be plotted for an ideal multiplier as shown in Figure 4-48a. Mismatches in emitter areas can be handled by defining an area mismatch factor γ .

$$\gamma = \frac{A_{E2}A_{E4}}{A_{E1}A_{E3}} \quad (4.77)$$

where A_{E1} is the emitter area of Q1 and so on. If $\gamma \neq 1$, the effect is an even-order distortion term added to the transfer characteristic as shown in Figure 4-48b. If parasitic base and/or emitter resistances are present, the effect is an odd-order distortion term as shown in Figure 4-48c.

The noise behaviour of the multiplier quad as a function of control voltage is also noteworthy. With the multiplier nulled, all four devices are biased in their active regions and all four contribute equally to the output noise current. As the multiplier is "steered" away from null, two of the devices conduct more heavily and two move towards cutoff. In the limit where two devices are completely cut off, they make no thermal or shot noise contributions to the output, and the loop of junctions is effectively broken. Therefore, the multiplier exhibits maximum output noise at null, with noise decreasing as the quad is steered away from null in either direction.

Figure 4-48 : Multiplier current gain for ideal devices (a), emitter area mismatch (b) and parasitic base resistance (c)

The circuit of Figure 4-49 is evaluated for output noise current at 140 MHz as the control voltage is swept from negative to positive; the results are shown in Figure 4-50. The maximum noise contribution is always at null, but the shape of the curve depends on the relative contributions of thermal and shot noise, which are determined by circuit temperature, device size and bias currents.

Figure 4-49 : Multiplier noise test circuit **Figure 4-50 : Output noise versus multiplier weight for the circuit of Figure 4-49**

4.12 Layout Considerations

If translinear and differential techniques are to be used effectively, extreme care must be used in circuit layout. Generally speaking, careful simulation will bring the design to a certain "optimum" state: any imperfections in the layout (avoidable or otherwise) will degrade the actual performance from the expected "optimum".

When possible, it is best to choose a line of electrical and thermal symmetry through the center of the die, and to lay out the circuit with as nearly perfect symmetry as possible on either side of this line. The basis for symmetry may be the complementary halves of a differential circuit, the I and Q axes of a quadrature communications circuit or the two channels of a dual-channel circuit. It is also advisable to lay out significant thermal sources along a single line near one edge of the die, perpendicular to the axis of symmetry as shown in Figure 4-51.

Figure 4-51 : Axis of symmetry and thermal source line for optimum die layout

The circuit will usually have one or more pairs of transistors which must be matched very closely. These devices should be on either side of the line of symmetry, and far from the thermal source line. Corners of the die should be especially avoided because

of the effect of mounting strain. This problem is caused by different thermal expansion coefficients between the die and the mounting eutectic or epoxy beneath it. The strain causes stress on the silicon crystal lattice, which in turn causes temperature-dependent piezoelectric mismatch voltages.

The best matching is obtained when the "cross-split" technique is used. Suppose differential pair Q1 and Q2 in Figure 4-52a are to be laid out for best matching. The first step is to split each device in two (each new device being half the size of the original). The devices are then connected diagonally as shown in Figure 4-52b. This configuration exhibits greatly reduced sensitivity to process, strain and thermal mismatches. Its disadvantage is the cumbersome wiring needed to make the connections to the emitter, base and collector of each device.

Figure 4-52 : "Cross-split" technique used to minimize thermal, strain and process offsets

Translinear circuits frequently make use of ratiometric design, where device emitter area ratios are intended to be integer multiples of each other. To do this with a 1:8 ratio, for example, the large device should have eight emitters each equal to the X1 emitter (in-

stead of one or two large emitters). Only in this way will the "edge effects" such as current crowding be exactly scaled for the two devices. In this situation it is also best to place the X1 device physically between two X4 devices connected in parallel electrically as shown in Figure 4-53. this layout will also average out thermal or strain gradients. In any circuit where matching is important, all emitter stripes must be oriented in the same direction on the die to maintain identical lithography and crystal orientation.

Figure 4-53 : Thermal "averaging" used in a ratiometric current multiplier

Most of the other layout issues arise from inherent imperfections in the metallic interconnect. In the SH3 process used for this work, the metal is 1 μm of gold, with a sheet resistance of 40 ± 20 milliohms per square. The temperature coefficient of resistance (TCR) is +2000 ppm/deg C for gold (+3900 ppm/deg C for aluminum), so any offset due to mismatched metal resistance will be highly temperature dependent. For this reason, it is essential that emitter metal be matched as shown in Figure 4-54. Base metal should also be matched wherever possible.

Figure 4-54 : Emitter metal matching

The current-carrying limit of metal in the SH3 process is 4 mA per micron of width; if this value is exceeded there is a risk of local "hotspot" fusing and electromigration. It is usually best to design metal runs for 1 mA per micron of width or less. It is important to remember that first- and second-layer metal are deposited in separate process steps, and therefore do not have correlated sheet resistances. Metal runs must be matched by metal on the same layer.

Power supplies usually disrupt the symmetry of the layout, and high-frequency integrated circuits should have two pins per power supply. The best power supply configuration uses two pins on opposite sides of the chip, and a "branching" topology as shown in Figure 4-55. Note that one metal layer is used for vertical runs, the other for horizontal runs.

Figure 4-55 : Proper use of dual bondpads and "branching" topology

Careful attention must be paid to parasitic resistance and capacitance of metal

runs. Inductance is much less of a concern because the metal runs are wide and thin. For example, a 10 mm long line of 4 μm width has a resistance of about 100 ohms, an inductance of about 10 nH and a capacitance to substrate of 0.7 pF. The inductive reactance of this line is equal to its resistance at 1.6 GHz; for shorter line lengths this frequency is even higher. As mentioned previously, resistance is calculated as 40 ± 20 milliohms per square. Capacitances are calculated using two components: an area component (pF/sq micron of overlap), and a fringing component (pF/micron of metal edge). For the SH3 process, the two components of metal-to-substrate capacitance for the two metal layers are:

Table 4.10: Metal-Substrate Parasitic Capacitance for SH3 Process

Layer	Area (pF/ μm^2)	Fringe (pF/ μm)
Metal 1	3.5×10^{-5}	5.5×10^{-5}
Metal 2	1.7×10^{-5}	4.5×10^{-5}

Using these guidelines, parasitic resistance and capacitance can be determined for any line section.

This leads naturally to the idea of post-layout simulation. If the parasitics are included for every node, simulation will be slow and unreliable because of the huge increase in node count and added "stiffness" (range of circuit time constants). Satisfactory results are obtained for current-mode circuits if resistance is included in series with high-current emitters and run capacitance is included at the cascode common-base collectors.

4.13 Applications

The most useful techniques for ultralinear wideband design are: the use of local feedback, the use of feedforward, appropriate choice of impedance levels, the use of current-mode techniques and symmetrical layout design. Not all these techniques are applicable to every circuit, so a designer should be aware of which techniques are appropriate

in a given situation.

The use of differential circuit topologies, symmetrical layout techniques and appropriate impedance levels is useful regardless of the bandwidth required. Besides the linearity benefits mentioned above, these techniques usually improve power supply rejection, as well as conducted and radiated electromagnetic interference.

The feedback techniques have some limitations on their useful bandwidth, which arise from the requirement for stability margins on closed-loop gain magnitude and phase. Simple resistive feedback can be applied up to the useful bandwidth of voltage-mode circuits (about $f_T/3$ for typical amplifiers). Resistive feedback may cause noise figure problems in low-noise front ends. The 3TB has a useful closed-loop bandwidth of about $f_T/5$: this is the maximum frequency for which it provides significant improvement to linearity. The small-signal bandwidth is more typically about $f_T/3$. Current feedback topologies also provide useful performance up to $f_T/5$.

In contrast to this, feedforward and current-mode techniques are useful at frequencies almost up to f_T . They usually demand a great deal of layout attention if maximum performance is to be achieved, but they are the best way to get wide bandwidth and high linearity without excessive power consumption. Although current-mode circuits are noisier than traditional circuits, they do provide a net increase in dynamic range.

It is also apparent that thermal distortion and device mismatches must be accounted for in simulation if high accuracy is needed.

5 SAW Driver Amplifier

5.1 Introduction

The research for this thesis was undertaken to develop a broad understanding of ultralinear wideband IC design. Several applications seemed attractive immediately, and one in particular required superlinear techniques to attain all the required performance specifications: a driver amplifier for surface acoustic wave (SAW) filters and delay lines.

The system under consideration is a modem for high-capacity digital microwave radio. The data rate is 311 Mbit/s, the symbol rate is 35.84 Mbaud and the channel bandwidth is 40 MHz. The modulation scheme is 512-state QAM, and the background error rate target is 10^{-13} (using error control coding). These demanding specifications must be maintained in the face of variable hardware imperfections and atmospheric conditions. The solution is to tightly control the channel filtering, and to use an adaptive equalizer to mop up linear degradations. In our implementation, a temperature-controlled SAW filter was used, and an adaptive transversal equalizer was implemented using a SAW delay line. The performance of a SAW device depends on the performance of the driver and pickup electronics used with it.

5.2 Design Requirements

A typical SAW device is shown in Figure 5-1. It is a piezoelectric crystal substrate with two or more transducers defined on its surface. The transducers are typically regularly-spaced interdigitated metal fingers. When an ac voltage is applied to the input transducer, the piezoelectric effect causes the crystal surface to be periodically stressed, and a propagating wave mode is supported near the surface of the crystal. When this acoustic

wave travels to the output transducer, a piezoelectric voltage is induced across the transducer and appears at the load. The interdigitated transducers are transversal structures, where the position of the transducer finger-pair determines the delay and the overlap of the fingers determines the "weight" assigned to that finger pair

Figure 5-1 : SAW Device

SAW devices have many characteristics which make them ideal for high-performance signal processing. They are physically small, cheap to manufacture and can provide performance unachievable with continuous-time filters, switched capacitor filters or digital signal processing (DSP). Since the transducers are defined lithographically, the filter accuracy and repeatability are determined by the metal deposition and etching. Because modern IC fabrication facilities offer excellent accuracy and repeatability, the filter characteristics are tightly controlled.

SAW devices do have some undesirable characteristics which must be taken into account. One problem is that the devices have high in-band loss; about 20 to 30 dB for the devices used in this work. The loss is due to the low efficiency of energy transfer in the electrical-acoustic-electrical operation of the device. Each transducer can be modelled as a three-port device, with two acoustic ports and one electrical port. Since the two acoustic ports are always matched, the electrical port must be mismatched with a consequently large return loss. This large loss must be compensated for by driving the device with a high-gain power amplifier to prevent unacceptable degradation of the system noise figure

(see Equation (2.25)). A second and related problem is that the terminal impedances of a SAW device are highly reactive and vary greatly with frequency within the passband. Circuit factor compensation [70] must therefore be used to design the SAW device, and assumptions must be made about the source and load impedances seen by it.

In addition, the SAW device is series-matched at both ports as shown in Figure 5-2. For the device used in this work, the impedance Z_1 must be about $60 + j100$ ohms and the impedance Z_2 must be about zero (the SAW device is modelled by a transconductance, with the output being an ac current).

SAW devices are also degraded by electromagnetic feedthrough (EMFT). Feedthrough is defined as any coupled or radiated signals which get from the input to the output transducer other than by acoustic propagation.

The transversal equalizer was implemented using a SAW delay line with one input transducer and 17 identical output transducers with time spacing equal to the symbol period. A driver amplifier is needed which can drive the matching network and SAW device at a high level while maintaining the stringent linearity specifications required by a high-order QAM system. The operating band is 120-160 MHz, nominal rms input level is -15 dBm and gain is 20 dB. The allowable magnitude and phase compressions at peak power are 0.1 dB and 1 degree, respectively.

The first implementation of the SAW driver amplifier was a two-stage differential topology, each stage being a common-emitter/common-collector signal splitter driving a totem pole output. The simplified schematic is shown in Figure 5-3.

Figure 5-3 : Simplified schematic of discrete SAW driver amplifier

Standard surface-mount components were used, including discrete transistors chosen for wide bandwidth and low noise. Two adjustments were included; an input amplitude balance control and a phase adjustment in one arm. The substrate was an eight-layer FR4 fiberglass circuit card measuring 60 mm by 80 mm. Total power dissipation was 2.5 W.

This first amplifier nominally worked, but there were some problems. The board area required was so large that the circuit had to be mounted on a daughter card above the IF signal processing section of the modem board. This led to increased expense in connectors, standoffs and a cable to connect the amplifier output to the matching circuit and SAW on the main board. This output cable radiated signal in a position-dependent manner because the grounds at each end of the cable could not be controlled properly. The balance and phase adjustments added expense to the assembly and test processes, and the power consumption was 100% over budget. Because the amplifier was physically distributed and had significant thermal gradients, "warmup drifts" occurred when the amplifiers were installed in the modem, and again when the modem was installed in the equipment bay. At moderate production volumes, the independent statistical variations of the discrete components caused yield problems. This led to expensive select-on-test specifications for these components.

Because of these problems, a feasibility study was started to evaluate IC solutions. IC amplifiers which met the electrical performance criteria were available, and would have removed the need for a daughter card with its associated problems. These designs still had excessive power dissipation and were even more costly than the discrete version. We therefore decided to see what performance could be achieved with an application-specific integrated circuit (ASIC), paying careful attention to the design considerations discussed in Chapter 4.

A loose constraint of 1 W power dissipation was established based on maximum ambient temperature in the unit, allowable maximum junction temperature and thermal properties of readily-available low cost packaging. The noise figure requirement for the amplifier was determined by the gains and noise figures of all preceding stages in the receiver chain (Equation (2.25)). Because the SAW driver is preceded by several high-gain, low-noise stages, the noise figure constraint was also loose; a maximum of 15 dB. The gain

was specified to be 19 to 21 dB. The third-order intermodulation levels (IM3) were specified at -65 dB. Maximum gain magnitude compression was set at 0.1 dB, and maximum phase compression was set at 1 degree.

The peak output signal level was determined by the rms input level, the gain and the peaking factors (Section 2.4). The input level is -15 dBm, or about 40 mV rms. With 20 dB gain, the output level is 400 mV rms. The constellation peaking factor for the circular 512-state QAM signal is 3 dB, and the filter peaking factor is 5 dB (for raised cosine shaping with $\alpha=0.116$ and $x/\sin(x)$ compensation). The peak output voltage is 8 dB above 400 mV, or 1 V peak.

In order to allow maximum flexibility for matching the SAW device, the amplifier was required to be a voltage-in, voltage-out device. The high-impedance input was shunt-matched to 50 ohms using an external resistor. The low-impedance output was series-matched with external resistors and an inductor. The output matching circuit was not completely determined at the time the amplifier was designed. It was decided to design the circuits on the Tektronix SH3 process based on high confidence in the process models. This constrained the choice of device models and focused the design effort on circuit evaluation.

5.3 SAW Driver Design

A differential cascode topology was chosen as shown in Figure 5-4. This circuit offers wide bandwidth, good linearity and moderate noise figure, as well as good reverse isolation. This last quality guarantees high, stable gain for a variety of output impedances.

Figure 5-4 : Differential cascode SAW driver amplifier topology

The simple circuit performed well, but the magnitude compression performance was inadequate. Because there was some margin available in the phase compression, Q1 and Q2 were replaced with 3TBs. This had the effect of greatly increasing the 0.1 dB compression point with only a small degradation in the 1 degree compression point. There was no improvement in the 1 dB compression point, but the system is unusable at this value of compression anyway. Further improvement was realized by replacing Q5 and Q6 with 3TBs as well.

The major problem in the amplifier design was the output current source/sink capability. In Figure 5-5, the output devices Q5 and Q6 source a peak current I_p in alternate directions through load resistor R_L . If R_L is 50 ohms and the peak load voltage is 1 V, I_p is 20 mA.

Figure 5-5 : SAW driver output current source/sink capability

In order that the “non-sourcing” transistor will not be cut off (for example, Q6 in Figure 5-5a), we must have the bias current $I_b > I_p$. This means that the peak current sourced from the emitter of Q5 or Q6 will be at least $2I_p$, or 40 mA. In order to carry this emitter current without distortion due to high-level injection, a large device must be used. This in turn causes bandwidth reduction and current drive problems in the preceding stage.

This problem was solved by replacing the fixed bias sinks I_b by dynamically-biased current sinks as shown in Figure 5-6. When Q5 is sourcing current I_L through the load, the sink with current mirror M2 is sinking current I_L (plus a small bias current for Q6). Mirror M1 is only sinking a small current, so Q5 only needs to drive the peak current I_p , or 20 mA.

Figure 5-6 : Dynamically-biased SAW driver output current sinks

While simple in principle, this output stage proved difficult to design in practice. The idle current in the output drivers and sinks tends to vary widely over the range of process variables and temperature, and the slew rate of the current sinks must be carefully maintained to keep intermodulation down. The determination of the effect of component mismatches is very time-consuming as well. After extensive simulation, it was concluded that a useful amplifier could be made in the SH3 process.

The circuit was designed as a Quickchip 2S analog array [71]. The array is essentially an ordered arrangement of transistors, resistors and capacitors surrounded by bondpads. The designer uses two metal layers, a via layer and a nichrome resistor layer to connect the pre-arranged devices. Compared to a full-custom IC design, the array has a lower design cost (about \$35K versus \$80K), and a faster prototype turnaround time (six weeks versus 14 weeks). In this case, fast turnaround was essential because the amplifier was replacing part of a subsystem that was already in production. The production price of

arrays is higher than that of custom chips because their layout is not optimized, but the application is in a low-volume product so this consideration is not important.

The design tools were Quickic and Tekspice; both were provided by Tektronix. Quickic is a graphic editor used for schematic capture, layout, layout-versus-schematic (LVS) and design rule checking (DRC). Tekspice is an enhanced SPICE simulator featuring superior convergence and graphics, nested variable sweeping and a wide variety of built-in models. An electrical rule checker (ERC) was used to verify that breakdown voltage and current density limits were not violated. These tools run on a Sun Sparcstation 2.

The package chosen for the chip was a 28-pin plastic leaded chip carrier (PLCC28) because of its cost, thermal properties and electrical properties. These are summarized in Table 5-1.

Table 5.1: PLCC28 Package Properties

Property	Value
Outside dimensions	12.4 X 12.4 X 4.4 mm
Min/Max die size	1.27/3.81 mm square
Thermal resistance, junction to ambient	85 deg C/W
Pin-to-pin capacitance	300 fF
Pin-to-substrate capacitance	<50 fF
Pin inductance	3 nH
Pin-to-pin mutual inductance	1.3 nH
Bondwire inductance	2 nH
Price	US\$4.25

The block diagram of the amplifier is shown in Figure 5-7 and the final schematic is shown in Figure 5-8.

Figure 5-7 : SAWD amplifier block diagram

Figure 5-8 : SAWD amplifier schematic

Power supplies and bias circuits

Input buffers and gain stage

Output driver and current sink (+)

Output driver and current sink (-)

Q1-Q2 and Q29-Q30 are bias generators for the current sinks on the chip. Bias is set at 2 mA per emitter, with 200 mV degeneration. BIAS1 is used for the input/gain stages and BIAS2 is used for the output stage; this prevents corruption of the low-level input by the high-current outputs. Compensation capacitors CC1-CC4 are used to tame the high-frequency response of the bias circuit. The basic cascode is made up of common-emitter pair Q9-Q16 and common-base pair Q25-Q28. Tail current for this amplifier is 16 mA, and its gain is supplemented by Q26-Q27 with a 2 mA tail current. Q26-Q27 also provide feed-forward compensation for Q25-Q28. The overall gain is $(18/16)(280/30) = 10.5$; this is all the gain since the output stage is essentially a voltage follower. Q9 and Q16 are included in 3TBs Q3-Q8 and Q17-Q22 respectively. The tail currents for the 3TBs are 4 mA each.

The output stage is also differential; consider the side that drives OUTP. The emitter follower Q37 is embedded in 3TB Q31-Q36. Tail current for this 3TB is 8 mA, reflecting the large current drive requirements of the output stage. The output mirror is Q53-Q55, and it is driven by a 3TB Q42-Q48 which imposes a voltage at the top of R54. Since the lower end of R54 is fixed at approximately two diode drops above ground, the 3TB/mirror combination is a voltage-controlled current source. Idle current is set to about 10 mA, and the maximum sink current is slightly greater than 20 mA as required. The OUTN side is identical.

Careful attention must be paid to power supplies and grounding in the design of such a chip. Four non-adjacent pins are used for each power supply (VCC at +5 V, VEE at -5 V). The supply distribution on-chip splits the input/gain stage supply from that for the output stage. Four pins are used for output power ground, two for the output stage buffer ground and two for the cascode common-base bias point.

The low-inductance output ground (separate from all other grounds) is very important in high-frequency design, and in this case it also provided some useful flexibility. The device met specifications as designed, but due to some additional loading on the out-

put, additional current source/sink capability was needed. If the output ground node is pulled below 0 V, the current mirror can sink more current (at the cost of higher idle current and power dissipation). This feature was used to provide increased margin in production.

The die floorplan is shown in Figure 5-9. The buffers B1 to B6 are those in Figure 5-7. The layout uses nearly perfect electrical and thermal symmetry. Note that the output devices are very close to the pads they are driving; this avoids degradations due to metal run resistance and unwanted signal coupling. The dominant thermal sources are laid out roughly in a line across the top of the die; the many ground leads in this area help to conduct heat out of the chip.

Figure 5-9 : SAWD die floorplan

5.4 SAW Driver Performance

The circuit (dubbed SAWD for SAW Driver) worked and met specifications on the first pass. Table 5-2 is a summary of the measured performance of the chip.

Table 5.2: Measured SAWD Performance

Parameter	Measured Value
3 dB bandwidth	700 MHz
Full-power bandwidth	300 MHz
Gain	20.4 dB
Reverse isolation	>60 dB
Noise figure	12 dB
IM3 at +5 dBm out	-68 dB
1 dB compression point (input)	+0.2 dBm
0.1 dB compression point (input)	-0.4 dBm
1 degree compression point (input)	-4.3 dBm
Power consumption	620 mW

Measured intermodulation was within 3 dB of simulation and measured compression points were within 1 dB of simulation.

The SAWD and associated components consumed a board area of 25 mm by 40 mm so the requirement for a daughter board was eliminated. The small physical size also made it easy to route signal and ground traces, so radiated and conducted EMFT levels were lower and more consistent from unit to unit. All adjustments were eliminated because of the tight matching available on-chip. Production price for the tested, packaged part was US\$7 in quantities of 1000 per year.

This performance was only achieved through careful attention to the design con-

siderations of Chapter 4. Table 5-3 shows the effect of various levels of detail in simulation. The “idealized” circuit is that shown in Figure 5-7, with ideal current sinks, buffers, mirrors and resistors. The “complete” circuit is that shown in Figure 5-8, with real device models substituted. The parasitics considered were the PLCC28 package parasitics and metal run resistance and metal run-to-substrate capacitance where appropriate. Thermal self-heating models were only used for Q9, Q16, Q25, Q28, Q37 and Q62. The simulated value of IM3 for various cases is quite sensitive to the modelling assumptions; in the table below simulated IM3 is for worst-case temperature and process (high temperature, low β).

Table 5.3: Effect of Modelling Assumptions on Simulated IM3

Assumptions	IM3
Idealized circuit	-85 dB
Complete circuit	-79 dB
Complete circuit, parasitics	-78 dB
Complete circuit, parasitics, thermal	-72 dB
Complete circuit, parasitics, thermal, one-sigma mismatches	-65 dB
Measured	-68 dB

In general, the sensitivity of simulated results to simulated degradations depends on process, operating conditions and circuit topology.

6 Conclusions and Further Work

The purpose of this work was to examine the system design aspects of distortion, the modelling considerations for predicting it and the circuit design techniques which reduce it. Concrete results were achieved in all these areas.

The system design considerations involve a suitable choice of distortion measure (intermodulation, compression or the like), and a proper treatment of noise and distortion simultaneously using the cascade relations (Equations (2.25) and (2.26)). For QAM systems, the system designers must determine the maximum allowable magnitude and phase compression from the modulation scheme, and the peak-to-rms power ratio from the constellation and filter impulse response. The system design phase offers the greatest potential for cost savings from intelligent partitioning of signal processing functions. This means that the performance and packaging requirements must be "solid" before detailed circuit design can begin.

Modelling and simulation can accurately predict distortion performance if proper attention is paid to certain factors which strongly influence this performance. For operating frequencies below $f_T/3$, the conventional GP2 SPICE model is usually adequate for bipolar transistors. The differential gain test and transient analysis are the most appropriate ways to model compression and intermodulation, respectively. Simulated and measured intermodulation agree to within 3 dB, and compression results agree to within 1 dB. Thermal models which account for dynamic self-heating of transistors were applied to increase accuracy in simulation. Thermal distortion should be accounted for at low frequencies and wherever very low intermodulation performance is specified. Although dynamic cross-heating was negligible in this work due to the large separations between devices, similar

models can be used to account for it if needed.

The area of simulation which needs the most development is statistical design. Modern processes are specified with known design targets for all process parameters, and allowable limits for deviation from the target values. A sensible calculation of yield can only be made if the means and standard deviations of these parameters are known, as well as any correlations between them. Monte Carlo design methods can be used to produce a statistical distribution of a performance measure, but the simulation resources required for this are often prohibitive. A better method might be to use models to calculate the sensitivity functions of the process random variables directly.

The design considerations which most strongly affect linearity are operating point, thermal distortion and device mismatches. Peak signal values must be known accurately to estimate compression and intermodulation. Intermodulation is very sensitive to slew-rate limiting, so the maximum value of the time derivative of the signal is also important. Intelligent consideration of the layout and packaging parasitics is necessary for maximum accuracy, especially near the upper frequency limit of device operation.

The recommended circuit techniques for ultralinear wideband integrated circuits include the use of local series and shunt feedback and three-transistor buffers (3TBs) with high open-loop gain. Translinear and current-mode circuits are very useful because they minimize the number of current-to-voltage and voltage-to-current conversions in the signal path. Any linear signal processing function (and many nonlinear ones) can be implemented in a "stacked" current-mode topology. The advantages of this topology are high linearity, wide bandwidth and efficient use of supply currents. The disadvantage is a large power supply headroom requirement.

An ultralinear wideband SAW driver amplifier (SAWD) was designed and fabricated in a 6 GHz bipolar process. It met specifications on the first pass and provided supe-

rior electrical performance to a discrete component amplifier. Compared to the discrete design, the IC implementation was 1/10 the cost, used 1/5 the area and 1/4 the power dissipation.

Several possibilities for further work exist. The SAWD has its 1 degree phase compression point 4 dB below the 0.1 dB compression point. Measurements performed on the demodulator showed that error rate was limited by phase compression, so there is considerable interest in improving the phase compression without degrading the other performance characteristics or increasing power consumption.

A newer Tektronix process (SHPi) is now available. The f_T is 9 GHz and parasitic capacitances and resistances are smaller than for SH3. In particular, the product of R_b and C_{jc} is about 1/7 that for SH3. In Chapter 4, this product was strongly identified with intermodulation and phase compression. It is therefore expected that a superior amplifier could be designed with the same power consumption.

One serious limitation to many high-speed processes is the poor quality of the lateral PNP transistors. In the 3TB, the PNP devices are not in the signal path, so their low speed is not a serious drawback. The major concerns are collector-base capacitance, low beta and a low knee current density for beta falloff. If the designer tries to reduce the current density by using multiple devices in parallel, the parasitic capacitance can become prohibitive. The best solution to the problem is a complementary process with fast vertical PNP and NPN devices. Besides superior 3TB performance, such a process also makes possible folded current-mode topologies and larger common-mode range for differential circuits. The combination of wide bandwidth, superb linearity and low power consumption will be essential for the next generation of personal communication systems.

The achievement of high linearity with low power consumption also makes it possible to package the SAW driver and pickup electronics with the SAW device in a multi-

chip module (MCM). The performance advantages possible with this approach include reduced coupling and feedthrough due to isolation from the circuit board, as well as the shielding effect of the MCM package itself. Because the physical length of the signal connections is small, higher interface impedances can be selected to help reduce power consumption. The design costs for the circuit card are reduced because all layout-sensitive nodes are enclosed within the package; the only external connections are power supplies and matched impedance input and output lines.

Many of the design considerations treated in this work are directly applicable to FET circuit design. Proper consideration of offsets is especially important for FET design because FETs tend to have poorer matching than bipolar devices. Some of the current-mode techniques can be applied to FET design, but the translinear principle does not apply.

Almost all of this work applies to heterojunction bipolar transistors (HBTs). In traditional homojunction bipolar transistors, a compromise must be made between high current gain with small base transit time (which demands a narrow, lightly-doped base), and low noise and intermodulation (which demands low R_B , and therefore a wider base with heavier doping). In a heterojunction device, a difference in bandgap energy between emitter and base materials is exploited to improve emitter injection efficiency without the need for light base doping. In addition, the high electron mobility of GaAs makes a narrow base less critical when this material is used. HBTs are available now with wider bandwidth and lower noise than either silicon BJTs or GaAs FETs. The combination of ultralinear design techniques with the superb dc and ac qualities of HBTs will set new performance standards for wideband ultralinear analog integrated circuits.

A Distortion Measures

To show the relation in principle between the various distortion measures, I have selected the simplest possible example to minimize algebraic churning: a memoryless amplifier with a linear gain term, and quadratic and cubic distortion terms. The output of the amplifier is expressed in terms of the input as:

$$y(x) = a_1x + a_2x^2 + a_3x^3 \quad (\text{A.1})$$

Usually we restrict the coefficients such that $a_1x \gg a_2x^2$, $a_1x \gg a_3x^3$; this is referred to as weak nonlinearity. Three cases of input excitation are applied in turn to the amplifier, and the output is calculated analytically for each case:

$$\text{Case 1 : } x(t) = A \cos(\omega t) \quad (\text{A.2})$$

$$\text{Case 2 : } x(t) = A (\cos(\omega_1 t) + \cos(\omega_2 t)) \quad (\text{A.3})$$

$$\text{Case 3 : } x(t) = A ((1 + m \cos(\omega_m t)) \cos(\omega_1 t) + \cos(\omega_2 t)) \quad (\text{A.4})$$

From Case 1 we can extract total harmonic distortion and compression information, from Case 2 we can examine intermodulation and from Case 3 we can derive cross modulation.

Case 1 :

$$x(t) = A \cos(\omega t) \quad (\text{A.5})$$

$$(x(t))^2 = \frac{1}{2}A^2 + \frac{1}{2}A^2 \cos(2\omega t) \quad (\text{A.6})$$

$$(x(t))^3 = \frac{3}{4}A^3 \cos(\omega t) + \frac{1}{4}A^3 \cos(3\omega t) \quad (\text{A.7})$$

Separating the output components by frequency:

$$y_0(t) = \frac{1}{2}a_2A^2 \quad (\text{A.8})$$

$$y_1(t) = \left(a_1A + \frac{3}{4}a_3A^3\right)\cos(\omega t) \quad (\text{A.9})$$

$$y_2(t) = \left(\frac{1}{2}a_2A^2\right)\cos(2\omega t) \quad (\text{A.10})$$

$$y_3(t) = \left(\frac{1}{4}a_3A^3\right)\cos(3\omega t) \quad (\text{A.11})$$

By the definition of total harmonic distortion in Chapter 2:

$$THD = \frac{\left(\frac{1}{2}a_2A^2\right)^2 + \left(\frac{1}{4}a_3A^3\right)^2}{\left(a_1A + \frac{3}{4}a_3A^3\right)^2} \quad (\text{A.12})$$

It is clearly seen from Equation (A.12) that THD will depend on the input level.

Compression can be examined with the aid of Equation (A.9). The ideal linear output of the amplifier has an amplitude a_1A and the actual output amplitude at the input frequency is $a_1A + 0.75a_3A^3$. The 1 dB compression point can be defined in terms of the input level as follows:

$$20\log\left(\frac{a_1A + \frac{3}{4}a_3A^3}{a_1A}\right) = -1 \quad (\text{A.13})$$

$$A^2 = -0.1450 \frac{a_1}{a_3} \quad (\text{A.14})$$

This expression gives the input level for a 1 dB compressed output. In this simple amplifier, a_1 and a_3 will have opposite signs and the output compresses. If a_1 and a_3 had the same sign, the amplifier would exhibit gain expansion and Equation (A.14) would have no real solution. Some amplifiers, especially class AB stages, do exhibit gain expansion in the region of the input/output characteristic near zero input amplitude. These amplifiers would have a describing equation including terms with powers higher than three, so Equation (A.9) would include terms of degree five, seven and so on. With this more complicated expression, regions of gain expansion and compression can be modelled.

Case 2:

$$x(t) = A (\cos(\omega_1 t) + \cos(\omega_2 t)) \quad (\text{A.15})$$

$$(x(t))^2 = A^2 + \frac{A^2}{2} \cos(2\omega_1 t) + \frac{A^2}{2} \cos(2\omega_2 t) + \quad (\text{A.16})$$

$$A^2 \cos((\omega_1 + \omega_2)t) + A^2 \cos((\omega_1 - \omega_2)t)$$

$$(x(t))^3 = \frac{9}{4} A^3 \cos(\omega_1 t) + \frac{9}{4} A^3 \cos(\omega_2 t) + \quad (\text{A.17})$$

$$\frac{A^3}{4} \cos(3\omega_1 t) + \frac{A^3}{4} \cos(3\omega_2 t) +$$

$$\frac{3}{4} A^3 \cos((2\omega_1 + \omega_2)t) + \frac{3}{4} A^3 \cos((2\omega_2 + \omega_1)t) +$$

$$\frac{3}{4} A^3 \cos((2\omega_1 - \omega_2)t) + \frac{3}{4} A^3 \cos((2\omega_2 - \omega_1)t)$$

Again we separate the output components by frequency:

$$y_0(t) = a_2 A^2 \quad (\text{A.18})$$

$$y_1(t) = \left(a_1 A + \frac{9}{4} a_3 A^3 \right) (\cos(\omega_1 t) + \cos(\omega_2 t)) \quad (\text{A.19})$$

$$y_2(t) = \frac{1}{2} a_2 A^2 (\cos(2\omega_1 t) + \cos(2\omega_2 t)) + \quad (\text{A.20})$$

$$a_2 A^2 (\cos((\omega_1 + \omega_2)t) + \cos((\omega_1 - \omega_2)t))$$

$$y_3(t) = \frac{1}{4} a_3 A^3 (\cos(3\omega_1 t) + \cos(3\omega_2 t)) + \quad (\text{A.21})$$

$$\frac{3}{4} a_3 A^3 (\cos((2\omega_1 + \omega_2)t) + \cos((2\omega_2 + \omega_1)t)) +$$

$$\frac{3}{4} a_3 A^3 (\cos((2\omega_1 - \omega_2)t) + \cos((2\omega_2 - \omega_1)t))$$

This information is sufficient to calculate the input power associated with second-order and third-order intercept points (IP2 and IP3 respectively):

$$\text{IP2: } a_2 A^2 = a_1 A \quad (\text{A.22})$$

$$A_{IP2}^2 = \left(\frac{a_1}{a_2} \right)^2 \quad (\text{A.23})$$

$$\text{IP3: } \frac{3}{4} a_3 A^3 = a_1 A \quad (\text{A.24})$$

$$A_{IP3}^2 = \frac{4a_1}{3a_3} \quad (\text{A.25})$$

From Equations (A.25) and (A.14) we note that third-order intercept input power is proportional to 1 dB compression point input power:

$$\frac{A^2(IP3)}{A^2(CP1dB)} = \frac{1.3333 \frac{a_1}{a_3}}{0.1450 \frac{a_1}{a_3}} = 9.1955 = 9.6 \text{ dB} \quad (\text{A.26})$$

This is the theoretical basis for the rule of thumb that an amplifier will be 1 dB compressed at an input power 10 dB below the input power for the theoretical third-order intercept point.

Case 3:

$$x(t) = A(1 + m \cos(\omega_m t)) \cos(\omega_1 t) + \cos(\omega_2 t) \quad (\text{A.27})$$

The expansion of $x^2(t)$ and $x^3(t)$ creates an enormous number of terms, most of which are irrelevant to cross-modulation. We are interested in the transfer of modulation from the signal at ω_1 to the carrier at ω_2 , in other words the components at $\omega_2 + \omega_m$ and $\omega_2 - \omega_m$. Only some of the terms in $x^2(t)$ are relevant to further consideration in the cubing operation:

$$(x(t))^2 = A^2 \cos((\omega_1 + \omega_2)t) + A^2 \cos((\omega_1 - \omega_2)t) + \quad (\text{A.28})$$

$$\frac{1}{2} A^2 m^2 \cos(\omega_m t) + \frac{1}{2} A^2 m^2 \cos(-\omega_m t) +$$

$$\frac{1}{2}m^2A^2 (\cos ((\omega_1 + \omega_2 + \omega_m)t) + \cos ((\omega_1 - \omega_2 + \omega_m)t)) +$$

$$\frac{1}{2}m^2A^2 (\cos ((\omega_1 + \omega_2 - \omega_m)t) + \cos ((\omega_1 - \omega_2 - \omega_m)t)) + \dots$$

We now calculate the $x^3(t)$ term which will contain some components near the original frequencies. If this term is expanded, retaining only those terms which contribute to cross-modulation products:

$$x^3(t) = \frac{3}{2}A^3m\cos((\omega_2 + \omega_m)t) + \frac{3}{2}A^3m\cos((\omega_2 - \omega_m)t) + \dots \quad (\text{A.29})$$

CM is the amplitude of the modulation components at $\omega_2 \pm \omega_m$:

$$CM = 3a_3mA^3 \quad (\text{A.30})$$

Recall from Equation (A.21) that the third-order intermodulation products have amplitudes:

$$IM3 = \frac{3}{4}a_3A^3 \quad (\text{A.31})$$

Thus when $m=1$ (100% modulation index), we obtain Equation (2.6):

$$\frac{CM}{IM3} = 4 \quad (\text{A.32})$$

Finally, the same amplifier characteristic can be treated for conversion linearity. We assume conversion by a linear quantizer (equidistant decision boundaries) of n bits, with the input signal range constrained to $(-X_{\max}, X_{\max})$. In Figure A-1, the ideal and actual gain characteristics are plotted along with the error function $\epsilon(x)$:

$$y_{ideal}(x) = a_1 x \quad (A.33)$$

$$y_{real}(x) = a_1 x + a_2 x^2 + a_3 x^3 \quad (A.34)$$

$$\varepsilon(x) = y_{real} - y_{ideal} = a_2 x^2 + a_3 x^3 \quad (A.35)$$

Figure A-1 : Ideal linear gain, real nonlinear gain and gain error for an amplifier

The main requirement on quantization is that the maximum magnitude of $\varepsilon(x)$ should be less than the distance between decision boundaries. Under the quantization conditions described above, there is a decision boundary at $y=0$ and the spacing between decision boundaries is:

$$\Delta y = \frac{a_1 X_{max}}{2^{n-1}} \quad (A.36)$$

The condition for correct quantization is therefore:

$$\max |\epsilon(x)| < \frac{a_1 X_{\max}}{2^{n-1}} \quad (\text{A.37})$$

Under the conditions where $a_1, a_2, a_3 > 0$ and $dy/dx > 0$ everywhere, the maximum of the error function occurs at $X_{\max}$, and if this maximum error just equals the decision interval where $x = X_{\max}$, the number of bits can be determined:

$$a_2 X_{\max}^2 + a_3 X_{\max}^3 = \frac{a_1 X_{\max}}{2^{n-1}} \quad (\text{A.38})$$

$$n = \text{int} \left(1 + \log \left(\frac{a_1}{a_2 X_{\max} + a_3 X_{\max}^2} \right) \right) \quad (\text{A.39})$$

where the logarithm has base two.

Linearity for wideband integrated circuits is traditionally quoted in "percent" or "dB" rather loosely, and this terminology should be firmly defined. Figure A-2 shows the relevant characteristic curves; $G_{\max}$ is the maximum value of small-signal voltage gain within the input range $(-V_m, V_m)$, and $G_{\min}$ is the minimum value of this gain.

Magnitude linearity in dB and in percent are defined as follows:

$$ML(\text{dB}) = 20 \log \left| \frac{G_{\max}}{G_{\min}} \right| \quad (\text{A.40})$$

$$ML(\%) = 100 \left(1 - \frac{G_{\min}}{G_{\max}} \right) \quad (\text{A.41})$$

Figure A-2 : V_{out} and dV_{out}/dV_{in} for obtaining G_{max} and G_{min}

B Distortion Limits for QAM Systems

We wish to consider transmission of a square QAM constellation, and the effect of magnitude-only and phase-only errors upon that transmission. Consider the series of diagrams in Figure B-1.

Figure B-1 : Peak state magnitude for various QAM constellations

The signal space has been normalized to ± 1 on both axes and we wish to calculate first the distance D to the peak (corner) state when the quantization is n bits on each axis. From the figure it is clear that:

$$D_n = \left(\frac{2^n - 1}{2^n} \right) \sqrt{2} \quad (\text{B.1})$$

Figure B-2 : Determination of maximum magnitude-only error (ϵ_m)

Now consider Figure B-2. The distance d represents the maximum magnitude-only error which, when added to the ideal vector, will not take the resultant outside the decision space. If the area of the decision space is A_{ds} and the length of one side is x_{ds} , then:

$$A_{ds} = \frac{\text{area}}{\text{points}} = \frac{4}{4^n} = 4^{1-n} \quad (\text{B.2})$$

$$x_{ds} = \sqrt{A_{ds}} = 4^{\frac{1-n}{2}} = 2^{1-n} \quad (\text{B.3})$$

Now the magnitude linearity error can be calculated as:

$$\varepsilon_m = \frac{d}{D} = \frac{2^{1-n} \cdot \frac{\sqrt{2}}{2}}{\left(\frac{2^n - 1}{2^n}\right)\sqrt{2}} = \frac{1}{2^n - 1} \quad (\text{B.4})$$

Figure B-3 : Determination of maximum phase-only error (ε_p)

In Figure B-3, the effect of a phase-only error is shown. The angle ε_p is the angular deviation which, when applied to the ideal signal vector, takes the resultant just to the perimeter of the decision space. For a corner state in a square constellation:

$$x_0^2 + y_0^2 = r^2 = 2x_0^2 = 2\left(\frac{2^n - 1}{2^n}\right)^2 \quad (\text{B.5})$$

For a vector deviated counterclockwise from (x_0, y_0) to (x_1, y_1) :

$$x_1 = x_0 - \frac{x_{ds}}{2} \quad (\text{B.6})$$

$$y_1^2 = r^2 - x_1^2 = x_0^2 + x_0 x_{ds} + \frac{1}{4} x_{ds}^2 \quad (\text{B.7})$$

The angles of the original and deviated vectors are:

$$\theta_0 = \text{atan}\left(\frac{y_0}{x_0}\right) = \text{atan}(1) = 45 \text{deg} \quad (\text{B.8})$$

$$\theta_1 = \text{atan}\left(\frac{y_1}{x_1}\right) = \text{atan}\left(\frac{\sqrt{x_0^2 + x_0 x_{ds} + \frac{1}{4} x_{ds}^2}}{x_0 - \frac{x_{ds}}{2}}\right) \quad (\text{B.9})$$

Using $x_0 = 1 - 2^n$ and $x_{ds} = 2^{1-n}$ in Equation (B.9), we obtain:

$$\theta_1 = \text{atan}\left(\frac{\sqrt{1 - 2^{1-2n}}}{1 - 2^{1-n}}\right) \quad (\text{B.10})$$

Finally, the phase linearity error in degrees can be calculated:

$$\epsilon_p = \theta_1 - \theta_0 = \text{atan}\left(\frac{\sqrt{1 - 2^{1-2n}}}{1 - 2^{1-n}}\right) - 45 \quad (\text{B.11})$$

The same method can be used to calculate ϵ_m and ϵ_p for non-square constellations, but the expressions will be much more complicated because the peak state magnitude will be a much more complex function of n . To get conservative estimates, the designer should assume the error bounds for the next largest square constellation. For example, the circular 512-state QAM constellation is a subset of the square 1024-state constellation, so the error bounds for 1024QAM can be assumed.

C QAM Constellation Peaking Factor

The peak-to-rms ratio for a square QAM constellation with n bits per axis is derived using the assumption that all states are equally probable. Consider a signal space normalized to ± 1 on each axis as shown in Figure C-1.

Figure C-1 : One quadrant of a square QAM constellation

We can limit discussion to the top right quadrant since the constellation has perfect symmetry. One quadrant contains 4^{n-1} states, and the peak state or corner state has co-ordinates $((2^n - 1)/2^n, (2^n - 1)/2^n)$. By definition:

$$\frac{peak}{rms} = \frac{|peak|}{\sqrt{\frac{\sum_{all} (|state|)^2}{4^n - 1}}} \quad (C.1)$$

The magnitude of the peak state is easily calculated:

$$|peak| = \sqrt{\left(\left(\frac{2^n - 1}{2^n}\right)^2 + \left(\frac{2^n - 1}{2^n}\right)^2\right)} = \sqrt{2} \left(\frac{2^n - 1}{2^n}\right) \quad (C.2)$$

Note that the states are numbered from bottom left to top right for summing. The sum of the squared magnitudes over all the states is:

$$\sum_{all} (|state|)^2 = \left(\left(\frac{1}{2^n}\right)^2 + \left(\frac{1}{2^n}\right)^2\right) + \quad (C.3)$$

$$\left(\left(\frac{3}{2^n}\right)^2 + \left(\frac{1}{2^n}\right)^2\right) + \left(\left(\frac{5}{2^n}\right)^2 + \left(\frac{1}{2^n}\right)^2\right) + \dots +$$

$$\left(\left(\frac{2^n - 1}{2^n}\right)^2 + \left(\frac{2^n - 1}{2^n}\right)^2\right)$$

It is apparent that the left-most (first) column and the bottom (first) row will together contribute 2^n terms of magnitude $(1/2^n)^2$ to the sum. The second row and second column will together contribute 2^n terms of $(3/2^n)^2$, and so on. The last row and last column will together contribute 2^n terms of $((2^n - 1)/2^n)^2$ to the sum. Thus the sum can be expressed as:

$$\sum_{all} (state)^2 = 2^n \cdot \sum_{i=1}^{2^n - 1} \left(\frac{2i - 1}{2^n}\right)^2 \quad (C.4)$$

Substituting Equations (C.2) and (C.4) into (C.1), we obtain:

$$\frac{pk}{rms} = \frac{\sqrt{2\left(\frac{2^n-1}{2^n}\right)^2}}{\sqrt{\frac{1}{4^{n-1}}\left(2^n \cdot \sum_{i=1}^{2^{n-1}} \left(\frac{2i-1}{2^n}\right)^2\right)}} = \sqrt{\frac{2^{3n-1} - 2^{n-1} + 2^{n-1}}{2^{n-1} \sum_{i=1}^{2^{n-1}} (2i-1)^2}} \quad (C.5)$$

In the case of non-square mappings (for example stepped-square or circular), the economy used in summing all the terms in Equation (C.3) will not be as great. It may be necessary or simpler to sum all the contributions point by point for a specific constellation. The definition of Equation (C.1) still applies, however. Note that the peak-to-rms ratio for each axis considered individually is the same as for the overall constellation as given in Equation (C.5).

D Noise and Intermodulation

The noise figure and n th-order intercept point of a cascade can be derived in terms of the noise figures and intercept points of the blocks in the cascade. Consider Figure D-1, a cascade of signal processing circuits, each with power gain, noise figure and n th order intercept point defined as shown. Gain and return loss are assumed constant over the band of interest.

Figure D-1 : Cascade of signal processing blocks

Each block "processes" both the input signal and input noise, and also adds noise of its own as shown in Figure D-2.

Figure D-2 : Input and output signal-to-noise ratio

Noise figure is defined as the ratio of input signal-to-noise ratio to output signal-to-noise ratio:

$$F = \frac{S_i/N_i}{S_o/N_o} = \frac{S_i N_o}{N_i S_o} \quad (\text{D.1})$$

The output signal power is just the input signal power multiplied by the product of the power gains:

$$S_o = S_i (G_1 G_2 \dots G_k) \quad (\text{D.2})$$

The output noise power from any given stage is multiplied by the power gain following that stage, so the total output noise power is:

$$N_o = (G_1 G_2 \dots G_k) N_i + (F_1 - 1) (G_1 G_2 \dots G_k) N_i + \\ (F_2 - 1) (G_2 \dots G_k) N_i + \dots + (F_k - 1) (G_k) N_i \quad (\text{D.3})$$

Substituting Equations (D.3) and (D.2) into (D.1) yields the well-known result:

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \dots + \frac{F_k - 1}{G_1 G_2 \dots G_{k-1}} \quad (\text{D.4})$$

Here F_i and G_i are unitless power ratios.

To handle intermodulation, we first obtain a relation between intermodulation product power, linear power and intercept point. Referring to Figure D-3:

Figure D-3 : Relation between P_{lin} , IP_n and P_{IMn}

$$\frac{n}{1} = \frac{(\Delta y_2) / (\Delta x)}{(\Delta y_1) / (\Delta x)} = \frac{IP_n - P_{IMn}}{IP_n - P_{lin}} \quad (D.5)$$

Solving Equation (D.5) for the intermodulation power P_{IMn} yields (all powers in dBm):

$$P_{IMn} = nP_{lin} + (1 - n)IP_n \quad (D.6)$$

Once again, all the blocks in Figure D-1 are assumed to have gain and return losses flat across the band of interest. A two-tone input to block 1 generates two output tones plus intermodulation products. Block 2 "amplifies" the tones and distortion, and adds distortion of its own in some phase relation to the distortion from block 1. The worst-case upper bound for intermodulation products at the output is calculated assuming that all the intermodulation contributions add in phase. Assuming the intermodulation power at the output is low compared to linear input power, the output power is:

$$P_{link} = P_{lin1} (G_1 G_2 \dots G_k) \quad (D.7)$$

The intermodulation voltage and power at the output are given by:

$$V_{IMn} = V_{IMnk} + V_{IMnk-1} \sqrt{G_k} + V_{IMn1} \sqrt{G_2 \dots G_k} \quad (D.8)$$

$$P_{IMn} = \left(\sqrt{P_{IMnk}} + \sqrt{P_{IMnk-1}} \sqrt{G_k} + \dots + \sqrt{P_{IMn1}} \sqrt{G_2 \dots G_k} \right)^2 \quad (D.9)$$

We now convert Equation (D.6) from dBm to watts, and substitute (D.7) and (D.9):

$$P_{IMn} = (P_{link})^n \cdot (IP_n)^{n-1} = (IP_n)^{n-1} \cdot P_{lin1}^n (G_2 \dots G_k)^n \quad (D.10)$$

$$(IP_n)^{\frac{1-n}{2}} = P_{lin1}^{\frac{-n}{2}} (G_2 \dots G_k)^{\frac{-n}{2}} \sqrt{P_{IMn}} \quad (D.11)$$

Next we obtain the intermodulation power at the output of each stage from Equation (D.10):

$$P_{IMn1} = (IP_{n1})^{1-n} \cdot P_{lin1}^n \quad (D.12)$$

$P_{IMn2} = (IPn_2)^{1-n} \cdot P_{lin1}^n (G_2)^n$, and so on up to

$$P_{IMnk} = (IPn_k)^{1-n} \cdot P_{lin1}^n (G_2 \dots G_k)^n$$

Now Equations (D.12) are substituted into Equation (D.11) to yield:

$$(IPn)^{\frac{1-n}{2}} = \frac{(IPn_1)^{\frac{1-n}{2}}}{(G_2 \dots G_k)^{\frac{n-1}{2}}} + \quad (D.13)$$

$$\frac{(IPn_2)^{\frac{1-n}{2}}}{(G_3 \dots G_k)^{\frac{n-1}{2}}} + \dots + (IPn_k)^{\frac{1-n}{2}}$$

Here IP_{in} are in watts and G_i are unitless power ratios.

E The SPICE GP2 Model

The most commonly used version of the SPICE modified Gummel-Poon model (GP2) is defined by a 15-element, 7-node model and its associated controlling equations. The user has a set of 39 physically-based device parameters available to define the behaviour of the model. These parameters fall into four basic groups, and will be defined first to simplify discussion of the model topology and controlling equations.

The first group of parameters is the set of geometry-independent parameters. A typical bipolar process offers one basic active device: the vertical NPN transistor. Since all devices are made simultaneously on the IC wafer, the junction depths and bulk region doping concentrations are highly consistent from device to device. Parameters which relate only to the "vertical" dimensions and properties of the device will therefore be the same for each device, regardless of its size. Device size is instead altered in the lateral dimensions. If the process is complementary (vertical PNP as well as NPN transistors), there will be one set of geometry-independent parameters for the NPN devices and another set for the PNPs. These parameters are as follows:

EG : Energy bandgap used for calculation of saturation current temperature dependence.

BF : Ideal (maximum) forward common-emitter current gain.

BR : Ideal (maximum) reverse common-emitter current gain.

VAF : Forward Early voltage.

VAR : Reverse Early voltage.

3 OF/DE 3

PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

NF : Forward current emission coefficient (associated with the "effective" component of base current, or diffusion current across the emitter -base junction).

NR : Reverse current emission coefficient.

NE : Base-emitter junction leakage emission coefficient (associated with the recombination component of base current for the emitter-base depletion region).

NC : Base-collector junction leakage emission coefficient.

VJE : Base-emitter depletion layer built-in voltage.

MJE : Base-emitter depletion layer exponential factor (or "grade coefficient").

VJC : Base-collector depletion layer built-in voltage.

MJC : Base-collector depletion layer exponential factor.

VJS : Collector-substrate depletion layer built-in voltage.

MJS : Collector-substrate depletion layer exponential factor.

XTF : Forward transit time collector-base bias dependence coefficient.

VTF : Forward transit time bias dependence exponent scaling voltage.

FC : Forward-bias depletion capacitance coefficient. This is needed because C_j does not increase as rapidly with forward bias as predicted by Equation (3.2).

Note that the parameters often appear in pairs corresponding to the forward and reverse modes. Forward mode is the normal operation condition of forward-biased emitter-base junction and reverse-biased collector-base junction. Reverse mode refers to the opposite bias conditions. Wideband linear ICs never use devices operated in reverse mode, so these parameters are mostly of secondary importance.

The second set of parameters are DC parameters. The values of these parameters are determined by both device geometry and geometry-independent processing factors. Together with the first set of parameters, the DC parameters completely define the DC behaviour of the device at fixed temperature. The DC parameters are:

IS : Transport saturation current. This parameter is used to scale collector current to $\exp(V_{BE}/v_t)$.

RB : Value of the base bulk resistance when collector current is zero.

RBM : Minimum value of bulk base resistance (observed at high collector currents).

IRB : Value of base current at which bulk base resistance is midway between RB and RBM.

RC : Bulk series collector resistance, including contact resistance.

RE : Bulk series emitter resistance, including contact resistance.

IKF : Forward current gain knee current. This is the collector current at which dependence of collector current on base-emitter voltage weakens to $I_C \propto \exp(V_{BE}/(2v_t))$.

IKR : Reverse current gain knee current.

ISE : Base-emitter leakage saturation current. This is associated with the recombination component of base current for the emitter-base depletion region.

ISC : Base-collector leakage saturation current.

Note that the DC base current has two terms: an "ideal" component that contributes to collector current:

$$I_{B1} = \frac{IS}{BF} \left(\exp \left(\frac{V_{be}}{V_t} \right) - 1 \right) \quad (E.1)$$

and a "non-ideal" component that contributes to recombination in the emitter-base depletion region:

$$I_{B2} = ISE \left(\exp \left(\frac{V_{be}}{NE \cdot V_t} \right) - 1 \right) \quad (E.2)$$

Once again, some of the parameters appear in forward/reverse mode pairs, with the forward mode being of the greatest interest.

The next set of parameters to be considered are the AC parameters. These parameters have no effect on the DC and bias analyses, but in conjunction with the preceding two sets of parameters they determine the large- and small-signal AC behaviour and noise properties. They are:

TF : Ideal forward base transit time. In an NPN transistor, this is the average time for an injected electron to cross the neutral base region from emitter depletion boundary to collector depletion boundary.

TR : Ideal reverse base transit time.

CJE : Base-emitter depletion region junction capacitance at zero bias.

CJC : Base-collector depletion region junction capacitance at zero bias.

CJS : Collector-substrate depletion region junction capacitance at zero bias.

XCJC : The fraction of the base-collector depletion capacitance connected to the internal base node. This means that (XCJC)(CJC) is connected to the internal base node and (1-XCJC)(CJC) is connected to the external base node.

ITF : Parameter for high-current effect on forward transit time. This is used to model the Kirk effect.

KF : Flicker noise (or 1/f noise) coefficient.

AF : Flicker noise exponent.

These last two parameters are used to model a base current source producing flicker noise:

$$i_{nbf}^2 = \frac{KF \cdot I_B^{AF}}{f} \quad (\text{E.3})$$

where I_B is the DC base current and f is frequency.

The final two parameters are used to control the temperature dependence of the model:

XTI : Exponent for temperature dependence of the transport saturation current.

XTB : Exponent for temperature dependence of forward and reverse common-emitter current gain.

Figure E-1 : Modified Gummel-Poon model for a bipolar transistor

The topology for the GP2 model is shown in Figure E-1. The charge control of the model is done through two constants representing the base width modulation (K_{b1}) and high-level injection (K_{b2}) effects:

$$K_{b1} = \frac{1}{1 - \frac{V_{be}}{VAR} - \frac{V_{bc}}{VAF}} \quad (E.4)$$

$$K_{b2} = \frac{(BF(T)) \cdot I_{be1}}{IKF} + \frac{(BR(T)) \cdot I_{bc1}}{IKR} \quad (E.5)$$

The two charge expressions are combined to produce a total normalized base charge:

$$K_b = \frac{K_{b1}}{2} \left(1 + \sqrt{1 + 4K_{b2}} \right) \quad (E.6)$$

The temperature-dependent current gains $BF(T)$ and $BR(T)$ are:

$$BF(T) = BF \cdot \left(\frac{T}{T_0}\right)^{XTB} \quad (E.7)$$

$$BR(T) = BR \cdot \left(\frac{T}{T_0}\right)^{XTB} \quad (E.8)$$

where T is the ambient temperature and T_0 is the nominal temperature (usually 300 K) where the model parameters are extracted (both temperatures in Kelvin). The transport saturation current $IS(T)$ attempts to model the temperature variation of IS :

$$IS(T) = IS \cdot \left(\frac{T}{T_0}\right)^{XTI} \exp\left(\left(\frac{EG}{v_t}\right)\left(\frac{T-T_0}{T_0}\right)\right) \quad (E.9)$$

The non-ideal (leakage) saturation currents are modelled similarly:

$$ISE(T) = ISE \cdot \left(\frac{T}{T_0}\right)^{\frac{XTI}{NE}} \left(\frac{T_0}{T}\right)^{XTB} \exp\left(\left(\frac{EG}{NE \cdot v_t}\right)\left(\frac{T-T_0}{T_0}\right)\right) \quad (E.10)$$

$$ISC(T) = ISC \cdot \left(\frac{T}{T_0}\right)^{\frac{XTI}{NC}} \left(\frac{T_0}{T}\right)^{XTB} \exp\left(\left(\frac{EG}{NC \cdot v_t}\right)\left(\frac{T-T_0}{T_0}\right)\right) \quad (E.11)$$

For a forward-biased base-emitter junction, the two components of base-emitter current are:

$$I_{be1} = \frac{1}{BF(T)} \left(IS(T) \left(\exp\left(\frac{V_{be}}{NF \cdot v_t}\right) - 1 \right) + GMIN \cdot V_{be} \right) \quad (E.12)$$

$$I_{be2} = ISE(T) \left(\exp\left(\frac{V_{be}}{NE \cdot v_t}\right) - 1 \right) \quad (E.13)$$

GMIN is a minimum "leakage" conductance used to help the solver converge to the correct solution; it is usually set at 10^{-12} mhos. For a reverse-biased collector-base junction, the two components of collector-base current are:

$$I_{bc1} = \frac{1}{BR(T)} (IS(T) (\exp(-5) - 1) + GMIN \cdot V_{bc}) + \quad (E.14)$$

$$\left(\frac{IS(T)}{NR \cdot v_t} \exp(-5) \right) \frac{(V_{bc} + 5 \cdot NR \cdot v_t)}{BR(T)}$$

$$I_{bc2} = ISC(T) (\exp(-5) - 1) + \quad (E.15)$$

$$\left(\frac{ISC(T)}{NC \cdot v_t} \exp(-5) \right) (V_{bc} + 5 \cdot NC \cdot v_t)$$

The collector-emitter current (or transport current) is controlled by total base charge K_b :

$$I_{ce} = \frac{BF(T) \cdot I_{be1} - BR(T) \cdot I_{bc1}}{K_b} \quad (E.16)$$

The intrinsic base current (excluding charge storage effects) is:

$$I_b = I_{be1} + I_{be2} + I_{bc1} + I_{bc2} \quad (E.17)$$

The intrinsic collector current (excluding charge storage effects) is:

$$I_c = I_{ce} - I_{bc1} - I_{bc2} \quad (E.18)$$

The base resistance R_{bb} depends on whether or not the base resistance change at high current levels is modelled. If not, then $IRB=0$ and:

$$R_{Bb} = RBM + \frac{RB - RBM}{K_b} \quad (E.19)$$

If $IRB > 0$ then:

$$R_{Bb} = RBM + (RB - RBM) \left(1 + 3 \left(\frac{I_b}{IRB} \right)^{0.852} \right)^{-1/2} \quad (E.20)$$

This completes the DC description of the model. The AC effects are described with the aid of a base transit time which varies with both base-collector bias and collector current:

$$TF(V, I) = TF + TF \cdot XTF \left(\exp \left(\frac{V_{bc}}{1.44 \cdot VTF} \right) \right). \quad (E.21)$$

$$\left(\frac{BF(T) \cdot I_{be1}}{BF(T) \cdot I_{be1} + ITF} \right)^2$$

The base-emitter diffusion capacitance is:

$$C_{be2} = \frac{\partial}{\partial V_{be}} \left(\frac{TF(V, I) \cdot BF(T) \cdot I_{be1}}{K_b} \right) \quad (E.22)$$

while the depletion region capacitance is:

$$C_{be1} = CJE \cdot (1 - FC)^{-MJE} \left(1 + MJE \frac{(V_{be} - FC \cdot VJE)}{VJE \cdot (1 - FC)} \right) \quad (E.23)$$

The additional component of base-emitter charge storage current is due to collector-base voltage modulation, and is defined by a capacitance:

$$C_{qbe} = \frac{\partial}{\partial V_{bc}} \left(\frac{TF(V, I) \cdot BF(T) \cdot I_{be1}}{K_b} \right) \quad (E.24)$$

Now the total base-emitter charge storage current is:

$$I_{becs} = (C_{be1} + C_{be2}) \frac{dV_{be}}{dt} + C_{qbe} \frac{dV_{bc}}{dt} \quad (E.25)$$

For the reverse-biased collector-base junction, the capacitances are the diffusion capacitance:

$$C_{bc2} = \frac{\partial}{\partial V_{bc}} \left(\frac{TR \cdot BR(T) \cdot I_{bc1}}{K_b} \right) \quad (E.26)$$

the internal depletion layer capacitance:

$$C_{bc1} = XCJC \cdot CJC \left(1 - \frac{V_{bc}}{VJC} \right)^{-MJC} \quad (E.27)$$

and the external depletion layer capacitance:

$$C_{bx} = (1 - XCJC) \cdot CJC \left(1 - \frac{V_{bc}}{VJC} \right)^{-MJC} \quad (E.28)$$

Finally, the collector-substrate capacitance is:

$$C_{cs} = CJS \left(1 - \frac{V_{cs}}{VJS} \right)^{-MJS} \quad (E.29)$$

F Functions of Random Variables

The effect of device mismatches on circuit performance can be explicitly treated by expressing the circuit function in terms of the device parameters, then expanding the circuit function using a hierarchical treatment of the device parameter statistics to find the statistics for the circuit function. This ultimately requires only a knowledge of the scaling, addition and multiplication properties of normally-distributed random variables.

If X and Y are independent random variables, and $E(X)$ is the mean of X and $V(X)$ is the variance of X , then the scaling, addition and multiplication properties of the random variables are:

$$E(a + bX) = a + bE(X) \quad (1)$$

$$V(a + bX) = b^2 V(X) \quad (2)$$

$$E(X + Y) = E(X) + E(Y) \quad (3)$$

$$V(X + Y) = V(X) + V(Y) \quad (4)$$

$$E(XY) = E(X) \cdot E(Y) \quad (5)$$

$$V(XY) = (E(X))^2 (E(Y))^2 \left(\frac{V(X)}{(E(X))^2} + \frac{V(Y)}{(E(Y))^2} \right) \quad (6)$$

$$E\left(\frac{X}{Y}\right) = \frac{E(X)}{E(Y)} \quad (7)$$

$$V\left(\frac{X}{Y}\right) = \frac{(E(X))^2}{(E(Y))^2} \left(\frac{V(X)}{(E(X))^2} + \frac{V(Y)}{(E(Y))^2} \right) \quad (8)$$

The practical difficulty with this approach is that many device parameters are correlated because different types of devices (for example, resistors and capacitors) are made simultaneously from the same few diffusions. One consequence of this is that a “worst-case” simulation in which all device parameters are set to their maximum possible mismatch or extreme values will be unnecessarily pessimistic.

The long-term solution to this difficulty is a set of physically-based device models which would include statistically variable lithographic and material properties. This might eventually reduce the enormous computing resources required for the full statistical design of a large analog IC.

References

- [1] D.C. Espley, "The Calculation of Harmonic Production in Thermionic Valves with Resistive Loads", Proc. IRE, October 1933, pp. 1439-1446.
- [2] E.L. Chaffee, "A Simplified Harmonic Analysis", Rev. Sci. Instr., October 1936, pp. 384-389.
- [3] T.T. Ha, Solid-State Microwave Amplifier Design, John Wiley and Sons, 1981, pp. 202-212.
- [4] J. Millman, C.C. Halkias, Integrated Electronics, McGraw-Hill, 1972, pp. 679-684.
- [5] P.R. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1977, pp. 286-290.
- [6] A.B. Grebene, Bipolar and MOS Analog Integrated Circuit Design, Wiley, 1984.
- [7] N.I. Meyer, "Nonlinear Distortion in Class A Amplifiers at Low and Medium Frequencies", Proc. IEE, May 1959, pp. 481-489.
- [8] H. Lotsch, "Third Order Distortion and Cross-Modulation in a Grounded Emitter Transistor Amplifier", IRE Trans. Audio, March 1961, pp. 49-56.
- [9] A.J. Malinckrodt, F.M. Gardner, "Distortion in Transistor Amplifiers", IRE Trans. ED, July 1963, pp. 288-289.
- [10] S. Narayanan, "Transistor Distortion Analysis Using Volterra Series Representation", BSTJ, May-June 1967, pp. 991-1024.
- [11] L.W. Nagel, "Efficient Computer Simulation of Distortion in Electronic Circuits", IEEE Trans. CT, November 1973, pp. 742-745.
- [12] J.J. Ebers, J.L. Moll, "Large-Signal Behaviour of Junction Transistors", Proc. IRE, pp. 1761-1769.
- [13] H.C. Poon, "Modelling of Bipolar Transistor Using Integral Charge-Control Model with Application to Third-Order Distortion Studies", IEEE Trans. ED, June 1972, pp. 719-731.
- [14] S. Narayanan, H.C. Poon, "An Analysis of Distortion in Bipolar Transistors Using Integral Charge Control Model and Volterra Series", IEEE Trans. CT, July 1973, pp. 341-351.
- [15] H.K. Gummel, H.C. Poon, "An Integral Charge Control Model of Bipolar Transistors", BSTJ, May-June 1970, pp. 827-852.
- [16] M.S. Nakhla, J. Vlach, "A Piecewise Harmonic Balance Technique for Determina-

- tion of Periodic Response of Nonlinear Systems", IEEE Trans CS, January 1976, p. 85.
- [17] A. Ushida, L.O. Chua, "Frequency-Domain Analysis of Nonlinear Circuits Driven by Multitone Signals, IEEE Trans. CS, September 1984, pp. 766-779.
 - [18] F.A. D'Altroy, R.M. Jacobs, J.M. Nacci, E.J. Panner, "Ultralinear Transistors", BSTJ, December 1974, pp. 2195-2202.
 - [19] H.E. Abraham, R.G. Meyer, "Transistor Design for Low Distortion at High Frequencies", IEEE Trans. ED, December 1976, pp. 1290-1297.
 - [20] H.C. Poon, "Implication of Transistor Frequency Dependence on Distortion", IEEE Trans. ED, January 1974, pp. 110-112.
 - [21] R.G. Meyer, R. Eschenbach, R. Chin, "A Wideband Ultralinear Amplifier from 3 to 300 MHz", IEEE JSSC, August 1974, pp. 167-175.
 - [22] K.H. Chan, R.G. Meyer, "A Low-Distortion Monolithic Wideband Amplifier", IEEE JSSC, December 1977, pp. 685-690.
 - [23] L.C. Thomas, "Eliminating Broadband Distortion in Transistor Amplifiers", BSTJ, March 1968, pp. 315-342.
 - [24] B. Gilbert, "A New Wideband Amplifier Technique", IEEE JSSC, December 1968, pp. 353-365.
 - [25] B. Gilbert, "A Precise Four-Quadrant Multiplier with Subnanosecond Response", IEEE JSSC, December 1968, pp. 365-373.
 - [26] W.M.C. Sansen, R.G. Meyer, "Distortion in Bipolar Transistor Variable Gain Amplifiers", IEEE JSSC, August 1973, pp. 275-282.
 - [27] T.T. Ha, Solid-State Microwave Amplifier Design, John Wiley and Sons, 1981, p. 202.
 - [28] J. Millman, C.C. Halkias, Integrated Electronics, McGraw-Hill, 1972, pp. 681-684.
 - [29] T.T. Ha, Solid-State Microwave Amplifier Design, John Wiley and Sons, 1981, p. 204.
 - [30] *ibid*, p. 207.
 - [31] R.G. Meyer, M.J. Shensa, R. Eschenbach, "Cross Modulation and Intermodulation in Amplifiers at High Frequencies", IEEE JSSC, February 1972, pp. 16-23.
 - [32] R.W. Lucky, J. Salz, E.J. Weldon Jr., Principles of Data Communication, McGraw-Hill, 1968, p. 42.
 - [33] F.G. Stremler, Introduction to Communication Systems (2e), Addison-Wesley, 1982, pp. 109-121.

- [34] *ibid*, pp. 368-370.
- [35] R.W. Lucky, J. Salz, E.J. Weldon Jr., Principles of Data Communication, McGraw-Hill, 1968, pp. 50-51.
- [36] J. Kimball, Tekspice User's Guide, Tektronix Inc., 1987.
- [37] H.K. Gummel, H.C. Poon, "An Integral Charge Control Model of Bipolar Transistors", BSTJ, May-June 1970, pp. 827-852.
- [38] I. Getreu, Modeling the Bipolar Transistor, Tektronix Inc., 1976, pp. 69-123.
- [39] *ibid*, pp. 43-68.
- [40] J.R. Hauser, "The Effects of Distributed Base Potential on Emitter-Current Injection Density and Effective Base Resistance for Stripe Transistor Geometries", IEEE Trans. ED, May 1964, pp. 238-242.
- [41] W.M. Webster, "On the Variation of Junction Transistor Current Amplification Factor with the Emitter Current", Proc. IRE, June 1954, pp. 914-920.
- [42] C.T. Kirk, "A Theory of Transistor Cutoff Frequency (f_T) Falloff at High Current Densities", IRE Trans. ED, March 1962, pp. 164-174.
- [43] J.M. Early, "Effects of Space-Charge Layer Widening in Junction Transistors", Proc. IRE, November 1952, pp. 1401-1406.
- [44] I. Getreu, Modeling the Bipolar Transistor, Tektronix Inc., 1976, p. 125.
- [45] Quickchip 2 Integrated Circuit Design Guide, Tektronix Inc., 1987, p. 4-6.
- [46] Tektronix Inc., Tekspice User Documentation, 1988.
- [47] *ibid*, p. 4-20.
- [48] S.W. Director, R.A. Rohrer, "Interreciprocity and its Implications", Proc. 1968 IEEE Int. Symp. Cct. Theory, September 1968.
- [49] A.V. Oppenheim, R.W. Schaffer, Digital Signal Processing, Prentice-Hall, 1975, p. 28.
- [50] W. Schaff, Digital Signal Processing: Techniques and Applications, Learning Tree International, 1990, p. 412-5-29.
- [51] F.J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", Proc. IEEE, January 1978.
- [52] *ibid*.
- [53] K.G. Schlotzhauer, A.J. Metz, U.S. Patent 4322688, March 30, 1982.
- [54] P. Quinn, "The Quickchip 2 Design Review", Design Guide Addendum, Tektronix

- Inc., May 1988, p. 4.
- [55] M.G. Bulmer, Principles of Statistics, Dover, 1979, pp. 68-80.
 - [56] R.C. Joy, E.S. Schlig, "Thermal Properties of Very Fast Transistors", IEEE Trans. ED, August 1970, pp. 586-594.
 - [57] K. Fukahori, P.R. Gray, "Computer Simulation of Integrated Circuits in the Presence of Electrothermal Interaction", IEEE JSSC, December 1976, pp. 834-846.
 - [58] G. Meijer, "The Current Dependency of the Output Conductance of Voltage-Driven Bipolar Transistors", IEEE JSSC, August 1977, pp. 428-429.
 - [59] J. Kuo, Y. Chen, "A Device-Level Transient Analysis of a 1 μ m Six-Transistor BiCMOS Inverter Circuit Using a Large-Scale Quasi-3D Device Simulator", Proc. IEEE BCTM 1991, pp. 93-96.
 - [60] R.M. Fox, S.G. Lee, "Predictive Modeling of Thermal Effects in BJTs", Proc. IEEE BCTM 1991, pp. 89-92.
 - [61] F.D. Waldhauer, Feedback, Wiley, 1982, pp. 295-323.
 - [62] D.J.H. Maclean, Broadband Feedback Amplifiers, Research Studies Press, 1982, pp. 123-144.
 - [63] A.B. Grebene, Bipolar and MOS Analog Integrated Circuit Design, Wiley, 1984, pp. 417-432.
 - [64] S. Simpkins, W. Gross, "Cascomp Feedforward Error Correction in High Speed Amplifier Design", IEEE JSSC, December 1983, pp. 762-764.
 - [65] J. Williams (Ed.), Analog Circuit Design, Butterworth-Heinemann, 1991, pp. 117-118.
 - [66] C. Toumazou, F.J. Lidgley, D.G. Hough, Analogue IC Design: the current mode approach, Peter Peregrinus, 1990, pp. 11-28.
 - [67] P.R. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1977, p. 618.
 - [68] C. Toumazou, F.J. Lidgley, D.G. Hough, Analogue IC Design: the current mode approach, Peter Peregrinus, 1990, pp. 53-63.
 - [69] B. Gilbert, "A New Wideband Amplifier Technique", IEEE JSSC, December 1968, pp. 353-365.
 - [70] T.P. Cameron, Circuit Factor Compensation for SAW Filters using Modal Analysis, M.Eng. Thesis, Carleton University, 1988.
 - [71] Quickchip 2 Integrated Circuit design Guide, Tektronix Inc., 1987, p. 1-5.

END

15-1 2-93

FIN