Nonlinear RF Circuits and Nonlinear Vector Network Analyzers

With increasingly low-cost and power-efficient RF electronics being demanded by today's wireless communication systems, it is essential to keep up to speed with new developments. This book presents key advances in the field that you need to know about and emerging patterns in large-signal measurement techniques, modeling, and nonlinear circuit design theory supported by practical examples.

Topics covered include:

- novel large-signal measurement techniques that have become available with the introduction of nonlinear vector network analyzers (NVNAs), such as the LSNA, PNA-X, and SWAP
- direct extraction of device models from large-signal RF dynamic loadlines
- characterization of memory effects (self-heating, traps) with pulsed RF measurements
- interactive design of power-efficient amplifiers (PAs) and oscillators using ultra-fast multi-harmonic active load-pull
- Volterra and poly-harmonic distortion (X-parameters) behavioral modeling
- oscillator phase noise theory
- balancing, modeling, and poly-harmonic linearization of broadband RFIC modulators
- development of a frequency-selective predistorter to linearize PAs

PATRICK ROBLIN is a Professor in the Department of Electrical and Computer Engineering at Ohio State University (OSU). He has worked at OSU since 1984, after receiving his D.Sc. degree in electrical engineering from Washington University. He is the founder of the Nonlinear RF Research Laboratory at OSU and previously co-wrote the book *High-Speed Heterostructure Devices* (Cambridge University Press, 2002).

THE CAMBRIDGE RF AND MICROWAVE ENGINEERING SERIES

Series Editor,
Steve C. Cripps, Distinguished Research Professor, Cardiff University

Peter Aaen, Jaime Plá and John Wood, *Modeling and Characterization of RF and Microwave Power FETs*
Dominique Schreurs, Máirtín O'Droma, Anthony A. Goacher and Michael Gadringer, *RF Amplifier Behavioral Modeling*
Fan Yang and Yahya Rahmat-Samii, *Electromagnetic Band Gap Structures in Antenna Engineering*
Enrico Rubiola, *Phase Noise and Frequency Stability in Oscillators*
Earl McCune, *Practical Digital Wireless Signals*
Stepan Lucyszyn, *Advanced RF MEMS*

Forthcoming
Matthias Rudolph, Christian Fager and David E. Root, *Nonlinear Transistor Model Parameter Extraction Techniques*
Sorin Voinigescu, *High-Frequency Integrated Circuits*
David E. Root, Jason Horn and Jan Verspecht, *X-Parameters*
John L. B. Walker, *Handbook of RF and Microwave Solid-State Power Amplifiers*
John Wood, Peter Aaen and Mohamed Sayed, *Modern RF Microwave Measurement Techniques*
Richard Carter, *Theory and Design of Microwave Tubes*
Anh-Vu H. Pham, Morgan J. Chen and Kunia Aihara, *LCP for Microwave Packages and Modules*
Nuno Borges Carvalho and Dominique Scheurs, *Microwave and Wireless Measurement Techniques*

Nonlinear RF Circuits and Nonlinear Vector Network Analyzers

Interactive Measurement and Design Techniques

PATRICK ROBLIN
Ohio State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521889957

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN 978 0 521 88995 7 Hardback

à Gloria

Contents

Preface

Introduction

This book is concerned with recent advances in the use of nonlinear vector network analyzers (NVNAs) for the characterization, modeling, design, and linearization of nonlinear RF devices and circuits.

Wireless has experienced an incredible growth; from cellular phones to wireless local area networks (WLANs), portable radios have become ubiquitous. The trend is to develop RF radios with wider bandwidth and low power dissipation for supporting new broadband services. Of particular importance are the nonlinear RF front-end circuits in radios, which typically dissipate half the handheld power. Given that this is a rapidly changing field, students, practicing engineers, and researchers must keep themselves abreast of new developments in the field. This book presents some of the emerging paradigms in large-signal measurement techniques, modeling, and nonlinear circuit design theory with supporting examples.

Brief description of the book

In Chapter 1, we will review the types of modulated RF signals involved in modern wireless communication, and introduce several measurement metrics used to characterize them. We will then discuss, in Chapter 2, novel measurement techniques that have become available with the introduction of NVNAs such as the large-signal network analyzer (LSNA) and, more recently, the PNA-X and SWAP. The NVNA permits the vector measurement of the fundamental and harmonics voltage and currents of devices or circuits for CW, modulated, and pulsed RF excitations.

Next, in Chapter 3, we will address the direct extraction of device models from the RF dynamic large-signal loadlines. Then, in Chapter 4, we show how memory effects (self-heating, traps) can be characterized with pulsed RF measurements.

Alternatively, bypassing modeling, we shall see, in Chapter 5, how to use nonlinear measurements for the direct interactive design of nonlinear RF circuits such as power amplifiers of various classes and oscillators, by using a new ultra-fast multi-harmonic active load-pull approach implemented with the sampler-based NVNA.

The behavioral modeling of nonlinear circuits using exact Volterra series, and poly-harmonic distortion (X-parameters[1]) and memory-polynomial approximations for both constant-wave and modulated signals will then be presented in Chapter 6.

The book will also show, in Chapter 7, how the output power and phase-noise properties of oscillators can be optimized using the measured device line. The Kurokawa phase-noise theory will be introduced both for free-running and for injection-locked oscillators and compared with alternative phase-noise theories.

Then, in Chapter 8, the balancing, poly-harmonic modeling and poly-harmonic predistortion linearization of RFIC modulators will be presented.

Finally, switching to the system level, we will see, in Chapter 9, how a frequency-selective predistorter can be developed to linearize PAs. An analytic technique for linearizing PAs modeled by memory polynomials will also be demonstrated for the two-tap case using a time-selective predistorter.

In summary, the selected examples presented in this book will provide an in-depth introduction to new emerging concepts and techniques for the large-signal measurement, modeling, and design of nonlinear RF circuits, while the literature referenced will allow the reader to further explore each topic as needed.

[1] "X-parameters" is a registered trademark of Agilent Technologies. The X-parameter format and underlying equations are open and documented.

"A great book for gurus as well as apprentices in the field of characterization and network analysis of nonlinear RF circuits. Prof. Patrick Roblin provides an extensive yet clear overview of this rapidly evolving field. The book enables RF engineers to connect the dots between a wide range of fascinating topics: modern modulation formats, NVNA and LSNA measurement technology, behavioral modeling, multitone measurements, power transistor memory effects, power amplifier distortion and predistortion, loadpull techniques, etc. By putting all of these topics together into one comprehensive book, Prof. Patrick Roblin proves that the whole field of network analysis of nonlinear RF circuits is much more than simply the sum of its parts."

Dr. Jan Verspecht, IEEE Fellow and inventor of X-parameters

"Since the advent of the nonlinear microwave CAD, in the 80s, RF nonlinear circuits were designed either recurring to this virtual lab environment, or to real but poor small-signal S-parameter measurements associated with incomplete amplitude-only power and spectrum characterization. It was only very recently that new instruments capable of large-signal RF characterization opened a new way to nonlinear RF circuit design. So, this book constitutes a unique and timely volume on the recent advances of nonlinear microwave circuit design based on these new time-domain waveform measurements that fills a gap faced by both industry engineers and academic researchers."

José Carlos Pedro, IEEE Fellow

Acknowledgments

This book on nonlinear RF measurements, modeling, and design would not have been possible without the help and contribution of many collaborators whom the author would like acknowledge here.

The author would first like to thank Professor Jean-Pierre Teyssier (XLIM, Brive, France), for his help and collaboration in the last decade, starting from the customization of the large-signal network analyzer (LSNA) system OSU acquired and extending to its recent upgrades to the SWAP operating software. Dr. Teyssier, who had designed the original mother board controlling the operation of the LSNA, kindly provided OSU with operating software upgrades and donated a pulse-controller daughter board for pulsed-RF measurements. The author would also like to thank Dr. Marc Vanden Bossche, and his collaborators Dr. Frans Verbeyst and Dr. Guillaume Paillocny of NMDG, for the outstanding technical support provided in the use of the LSNA.

The author is particularly indebted to his M.S. and Ph.D. students for their dedicated work in the laboratory. The early LSNA users at OSU included Suk Keun Myoung, Seok Joo Doo, and Xian Cui. The first LSNA measurements at OSU were performed by Dr. Suk Keun Myoung and Dr. Chaillot, who synchronized an FPGA testbed with the LSNA. A large fraction of the LSNA measurements reported in this book was conducted by Professor Seok Joo Doo (Korea Army Academy, South Korea), while he was a Ph.D. student at OSU. His talent, dedication, and thoroughness has greatly served the nonlinear RF research at OSU. Professor Doo also helped train new LSNA operators, Youngseo Ko and Chieh Kai Yang. Dr. Xian Cui (Skyworks Solutions, Nokia) while a Ph.D. student at OSU worked on the implementation of the multi-harmonic real-time active load-pull and its application to power amplifier design.

This book features the published work of several of the author's past Ph.D. students, namely Dr. Wenhua Dai (Avago Technologies), Professor Jayanta Mukherjee (India Institute of Technology, Bombay), Professor Seok Joo Doo (Korea Army Academy, South Korea), Dr. Xian Cui (Skyworks Solutions, Nokia), and Dr. Suk Keun Myoung (Samsung Corporation); his past-postdoc Dr. Fabien De Groote (VTD); his past M.S. students Mr. Venkatesh Balasubramanian and Mr. Shashank Mutha; and his present Ph.D. graduate students Inwon Suh, Chieh Kai Yang, Xi Yang, Young Seo Ko, and Ji Woo Kim. Unpublished material and material reported in theses and dissertations are credited in the book.

The author is particularly indebted to Dr. Dominique Chaillot, who spent two separate years in the OSU Nonlinear RF laboratory, starting in 2003 and 2007, respectively, and

trained two generations of OSU students in the use of FPGAs. In this book the author has credited his seminal research contribution by the naming (in Chapter 6) of the Chaillot functions, a useful expansion he suggested for the linearization of amplifiers.

The author is greatly appreciative of Professor Wan Rone Liou (National Taipei University, Taiwan) for his collaboration on several TSMC RFIC chips for the design of power RF amplifiers. The author would like to thank Dr. John Wood of Freescale Semiconductor for providing OSU with LDMOSFET power amplifiers used in the linearization, and Dr. Ray Pengelly of Cree, Inc. for providing OSU with GaN HEMT Doherty power amplifiers used in broadband measurements and behavioral modeling.

The author is grateful to his present and former students and collaborators in his group for their reading of the final manuscript and suggestions for improvements. He has acknowledged their multiple contributions in the appropriate chapters. However, the author takes full responsibility for the validity of the material presented. Note also that some of the material presented in this book includes new theoretical developments, modeling and linearization techniques, and measurement results, for which this book is the original reference.

The conduct of research is hardly possible without funding. The original acquisition of the LSNA was funded by an Air Force DURIP grant supervised by Dr. Gerald Witt with cost sharing provided by the State of Ohio and the Ohio State University (OSU). The largest portion of the work reported in this book was supported by a three-year GOALI grant[1] from the US National Science Foundation (NSF) in the ECCS (Electrical, Communications and Cyber Systems) division. An additional NSF grant[2] as part of the OSU Connection One IUCRC provided some of the support for the work on behavioral modeling and mixer linearization. The initial research on trap characterization in GaN HEMTs was funded by a contract from Wright Patterson Air Force Research Laboratory organized by Dr. Gregg Jessen. The model verification work on the 65-nm MOSFET was supported by Mr. Richard Taylor and Dr. Siraj Akhtar of Texas Instruments. Collaboration as part of this grant with Dr. Krishnanshu Dandu at Texas Instruments is also acknowledged. The research on combined LSNA and DLOS (deep-level optical spectroscopy) measurements for studying traps in GaN HEMTs was supported by a grant[3] from the Office of Naval Research organized by Dr. Paul Maki. Collaboration with Professor Ringle and Dr. Aaron Arehart of OSU for setting up the DLOS system is gratefully acknowledged. Research on linearization and behavioral modeling was supported by a grant from Samsung Cooporation procured by Dr. Myoung. The author is also particular grateful to Mr. Jeffrey Strahler and Andrew Coorporation for the support of the OSU Nonlinear RF lab with the loan of several key pieces of RF test equipment. The author would also like to thank Maury Corporation for loaning OSU a tuner and the Altera University program for donating several FPGA boards.

Research in a university is not possible without a collegial environment. The author is particularly grateful to his colleagues at OSU, Professors Stephen Ringel, John Volakis,

[1] ECS-0622003.
[2] IIP-0631286.
[3] N0014-08-1-0101.

Roberto Rojas, Ismail El-Nagar, Steve Bibyk, Siddarth Rajan, Wu Lu, Betty Lise Anderson, Furrukh Khan, JoAnne DeGroat, Paul Berger, George Valco, and Waleed Khalil, for their collaboration in some of the research projects and/or participation in theses committees. The author also greatly benefited from discussions and collaboration with visiting Professor Young Gi Kim (Hanyang University, South Korea), Hyo Dal Parker (Inha University, South Korea) and Zhijian Xie (North Carolina Agriculture and Technical State University).

The author is grateful for the welcome and support provided by the ARFTG (automatic RF technical group), particularly Dr. Jan Verspecht, Dr. Kate Remley, Dr. David Root, Professor Dominique Schreurs, Professor Jean-Pierre Teyssier, Dr. Dylan Williams, and Dr. John Wood. In particular the author is grateful for an educational student grant awarded to Dr. Seok Joo Doo by the ARFTG under the supervision of Mr. David Walker. The ARFTG conferences and the NVNA users' group provide an outstanding forum under the IEEE umbrella for the dissemination and promotion of research on linear and nonlinear RF measurements.

Finally, the author would like to thank Dr. Steve Cripps for his participation in the exciting Cambridge RF and Microwave Engineering Series he is editing. The author would like to express his gratitude to Dr. Julie Lancashire, Cambridge University Press publisher, for initiating this project and for her insight, guidance, and encouragement throughout the book's development. Special thanks goes also to Ms. Sarah Matthews for her highly professional help and support throughout this process.

Last, but not least, the author is grateful to Dr. Steven Holt for his careful copyediting of the manuscript.

1 Wireless signals[1]

Before we embark on a discussion of the characterization and modeling of nonlinear RF devices and circuits, let us review the attributes of their customers, the RF signals used in wireless communication. In this chapter, we will briefly discuss the trends of wireless communication systems and standards for cellular networks, wireless local area networks (WLANs), and wireless metropolitan networks (WMANs) and the challenges they create in the development of low-cost, linear, and power-efficient RF electronics. Various metrics used to characterize signals and systems will then be introduced.

1.1 Modern wireless communications

Wireless communication systems rely on the modulation of radio-frequency signals to exchange information. In modern wireless communication, digital modulation techniques whereby the digital information is encoded using amplitude (ASK), phase (PSK) or frequency (FSK) shift keying are used. In the most general case, quadrature amplitude modulation (QAM), whereby the amplitude $A(t)$ and phase $\phi(t)$ of the RF waves are both modulated, is used:

$$\begin{aligned} x_{\mathrm{RF}}(t) &= A(t)\cos[\omega_{\mathrm{RF}}t + \phi(t)] \\ &= I(t)\cos(\omega_{\mathrm{RF}}t) - Q(t)\sin(\omega_{\mathrm{RF}}t). \end{aligned}$$

In the above equation, $I(t)$ and $Q(t)$ represent the in-phase and quadrature components of the modulation, respectively.

It is to be noted that constant-envelope signals ($I^2(t) + Q^2(t) = \text{constant}$) are advantageous for mobile transmitters because they can then be amplified by nonlinear power amplifiers, without AM-to-PM distortions degrading the modulation of the signals transmitted. Such nonlinear amplifiers have the advantage of being both cheaper and more power-efficient than linear amplifiers.

Phase modulation is then attractive if it can be implemented in such a way that the phase-modulated signals can maintain their constant envelope $A^2 = I^2(t) + Q^2(t)$ while being processed by bandpass-limited circuits. One modulation scheme, GMSK (Gaussian minimum shift keying) [1], achieves this ability via the use of gradual

[1] Research collaboration with Jiwoo Kim and Suk Keun Myoung is gratefully acknowledged.

phase transition, which greatly helps maintain the signal envelope constant. The quasi-constant envelope of GMSK is beneficially used in wireless standards such as the GSM (Global System for Mobile Communication). The resulting low cost and power efficiency of the mobiles certainly contributed to the widespread success of this protocol around the world.

Nevertheless, given the needs to increase the wireless network's capacity to handle more users and to provide wide-bandwidth data services, digital quadrature amplitude modulation is becoming more prevalently used in modern standards.

Beside the modulation of the RF signals to encode information, the multiple division access (MDA) scheme used to handle multiple users or to increase the bandwidth for a single user is also of great importance. As symbolically represented in Figure 1.1, frequency (FDMA), time (TDMA), and code (CDMA) division multiple-access techniques are typically used to handle multiple users [1]. In CDMA, each user baseband signal is further modulated by a pseudonoise code with high chip-rate which spreads the bandwidth of the original baseband signal. Clearly CDMA calls for mobiles with front ends operating with wider frequency bandwidth.

Note that often the uplink and downlink between the mobile and the basestations use different frequencies, therefore relying on frequency-division duplexing. Alternatively, time-division duplexing (transmit and receive at different times) is used when the same frequency band is used.

In addition, the concept of the cellular network provides a frequency reuse in non-adjacent cells, which increases the capacity. Space-division multiplexing access (SDMA), or the use of antennas with beam-forming techniques, also enables further increase of the user capacity by adding additional sectors in each cell.

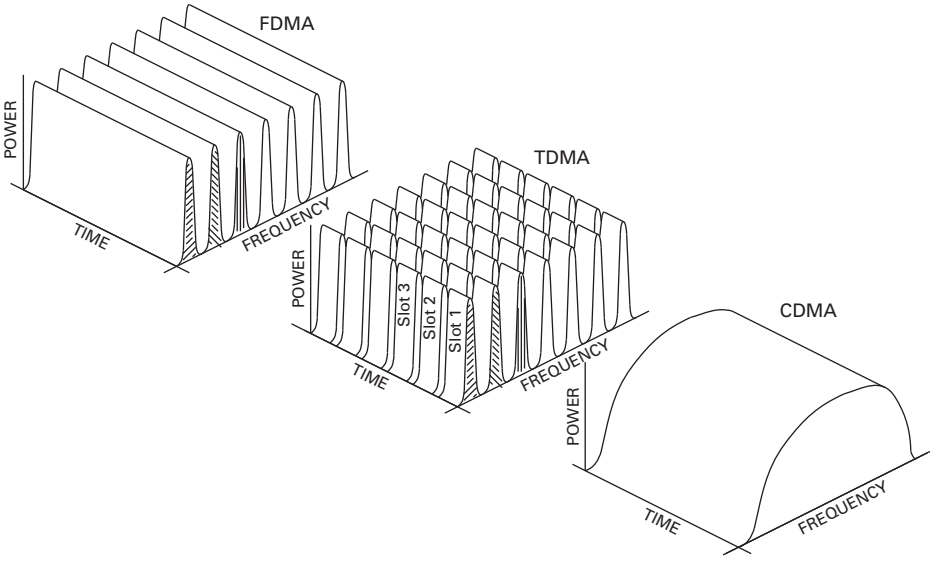


Figure 1.1 Comparison of FDMA, TDMA, and CDMA schemes.

Unlike the mobiles which in TDMA and FDMA handle a single channel, the power amplifiers at the basestation transmitter must typically operate with much wider bandwidth because it is more economical to use a single basestation RF power amplifier for the various channels for each RF carrier. Furthermore, basestation power amplifiers typically handle multiple carriers at the same time. This obviously introduces some critical constraints upon the linearity and bandwidth of these power amplifiers.

One of the motivations for the introduction of CDMA was its improved performance in handling fading (signal cancellation arising from multi-path propagation), owing to its wider bandwidth. CDMA can also make beneficial use of the various multi-path signals using a "rake receiver" [1] to reduce the signal-to-noise ratio at the output. Another division multiplexing technique that has gained prevalence in wireless communication systems due to its high performance in multi-path environments is OFDMA (orthogonal frequency-division multiplex access). OFDMA is a multi-user version of OFDM (orthogonal frequency-division multiplexing), which is used in high-performance WiFi WLANs, and is now being actively implemented in new WMAN standards such as WiMAX and LTE to provide wide-bandwidth data services [2] [3] [4]. Given the rapid growth in this field, we will discuss the characteristics of OFDM signals in the next section. The analysis of OFDM signals will also illustrate the challenges modern modulation schemes place upon the characterization and design of modern RF circuits.

1.2 OFDM primer

In this section we shall give a brief review of the most salient features of OFDM. OFDM signals typically consist of multiple symbols separated by a guard interval. The general structure of an idealized OFDM symbol of duration T_{SYMB} is shown in Figure 1.2, where $W_{T,\mathrm{SYMB}}(t)$ is an ideal shaping window.

The symbol duration is given by $T_{\mathrm{SYMB}} = T_{\mathrm{IFFT}} + T_{\mathrm{GI}} + T_{\mathrm{TR}}$ with T_{GI} the guard interval and T_{TR} the transition interval. The guard interval T_{GI} and the associated cyclic prefix $T_{\mathrm{CP}} = T_{\mathrm{GI}} + T_{\mathrm{TR}}$ are used to remove the intersymbol interference (ISI) from symbol to symbol as well as to maintain the subcarrier orthogonality in each symbol as we shall discuss below. The transition interval T_{TR}, which is not always specified in

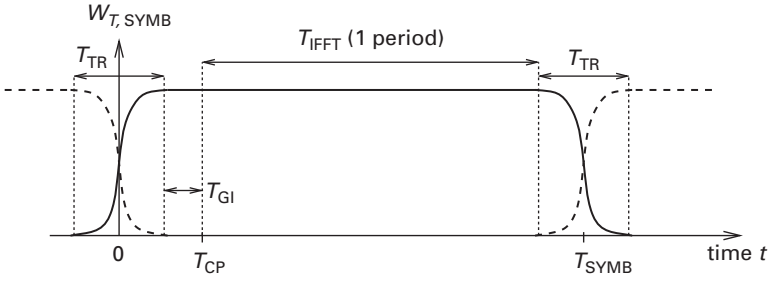


Figure 1.2 Structure of an OFDM symbol of duration T_{SYMB}.

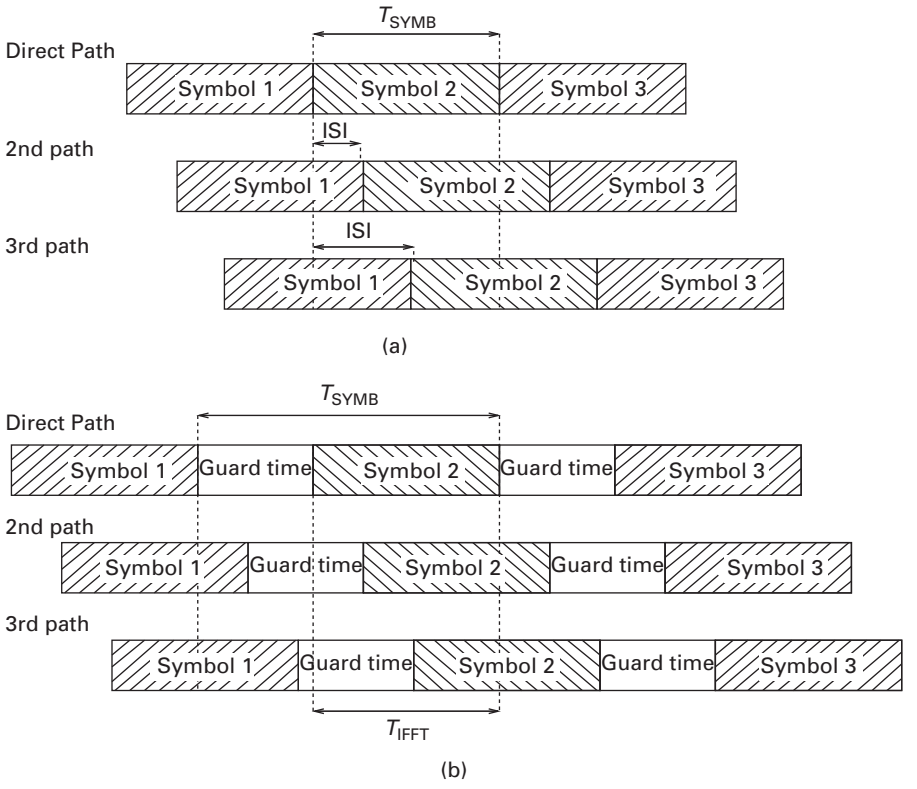


Figure 1.3 The ISI shown to take place in (a) is suppressed in (b) by the use of a guard interval between symbols.

the standards, is used to reduce the spectral regrowth associated with the transition from one to another symbol.

The use of the guard interval is illustrated in Figure 1.3. The intersymbol interference is suppressed if the guard time is larger than the channel delay spread. However, self-interference in a given symbol still needs mitigation, and this will be discussed later on.

Let us consider now a single OFDM symbol with index k. The data portion $x_{\mathrm{RF},k}$ which is located in the interval $kT_{\mathrm{SYMB}} + [T_{\mathrm{CP}},\ T_{\mathrm{CP}} + T_{\mathrm{IFFT}}]$ is of the form

$$x_{\mathrm{RF}}(t) = \mathrm{Re}\left\{\sum_k x_k(t - kT_{\mathrm{SYMB}})\exp\left[j2\pi f_0 t\right]\right\},$$

$$x_k(t) = \mathrm{rect}\left(\frac{t - (T_{\mathrm{CP}} + T_{\mathrm{IFFT}}/2)}{T_{\mathrm{IFFT}}}\right)\sum_{n=1}^{N_{\mathrm{SC}}} C_{k,n}\exp\left[j2\pi n\ \Delta f(t - T_{\mathrm{CP}})\right].$$

The Fourier transform $X_k(f) = \mathcal{F}[x_k]$ associated with the symbol x_k is

$$X_k(f) = T_{\mathrm{IFFT}}\sum_{n=1}^{N_{\mathrm{SC}}} C_{k,n}\exp\left[-j2\pi n\ \Delta f(2T_{\mathrm{CP}} + t_{\mathrm{IFFT}}/2)\right] \times \mathrm{sinc}\left(\frac{f - n\,\Delta f}{\Delta f}\right).$$

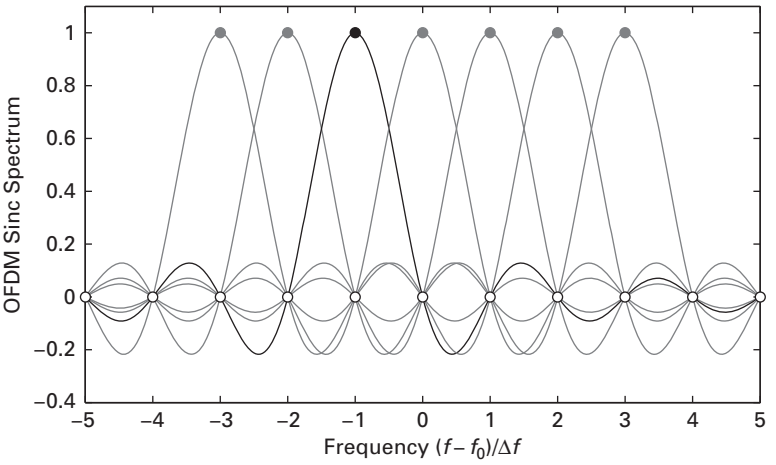


Figure 1.4 Normalized spectrum for seven adjacent subcarriers of an OFDM signal with subcarrier # −1 highlighted.

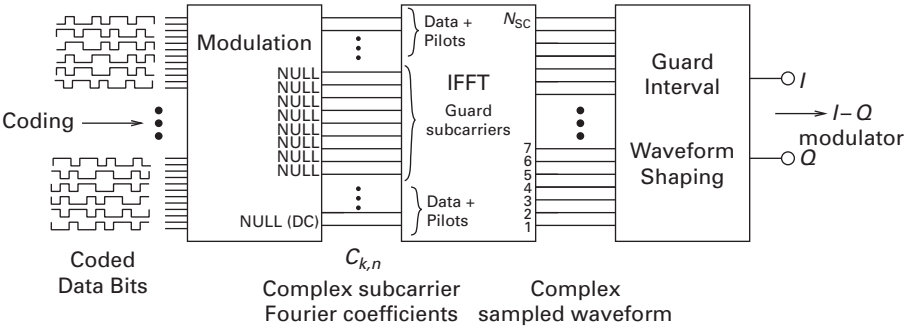


Figure 1.5 Transmitter implementation using IFFT.

The spectrum of the OFDM signal is obtained from the coherent superposition of the individual spectra of the various subcarriers. As can be seen in Figure 1.4, the spectra of the various subcarriers overlap in most of the frequency range. However, if we sample the frequency at $n\,\Delta f = n/T_{\mathrm{IFFT}}$, the spectra are orthogonal. This frequency sampling is equivalent to assuming that the signal $x_k(t)$ is a periodic function of time t with period T_{IFFT}. The orthogonality in OFDM is then achieved by selecting $T_{\mathrm{IFFT}} = 1/\Delta f$. The coefficients $C_{k,n}$ can then be recovered by performing a Fourier series over the data interval $kT_{\mathrm{SYMB}} + [T_{\mathrm{CP}},\ T_{\mathrm{CP}} + T_{\mathrm{IFFT}}]$. For sampled data, a Fourier series is exactly implemented (up to the scaling factor N_{IFFT}) by a fast Fourier transform (FFT) performed using N_{SC} data acquired during the data interval. Conversely, the sampled signal $x_k(nT_{\mathrm{s}})$ can be generated by an inverse FFT (IFFT) as indicated in Figure 1.5.

The complex Fourier coefficients C_k are typically generated using BPSK, QPSK, 16-QAM or 64-QAM modulation:

$$C_{k,n} = (I_{k,n} + j\,Q_{k,n}) \times K_{\mathrm{MOD},k},$$

Table 1.1. Parameters for WiFi, WiMAX (downlink), and LTE (downlink) for their widest-bandwidth configuration. Note that different circular prefixes (normal/extended) are used in LTE depending on the channel delay spread and the environment (urban/rural).

	WiFi 802.11a	WiMAX 806.16e-2005	LTE
Bandwidth	20 MHz	20 MHz	20 MHz
Number of subcarriers (N_{SC})	64	2048	2048
Number of data subcarriers	48	1440	1200
Number of null subcarriers	12	368	
Number of pilot subcarriers	4	240	
Subcarrier frequency (Δf)	0.3125 MHz	10.94 kHz	15 kHz
IFFT period ($T_{IFFT} = 1/\Delta f$)	3.2 μs	91.41 μs	66.67 μs
Circular prefix (T_{CP})	0.8 μs	11.4 μs	5/16.67 μs
Ratio T_{CP}/T_{IFFT}	1/4	1/8	1/14–1/4
Symbol interval (T_{SYMB})	4.0 μs	102.9 μs	71.67/183.34 μs

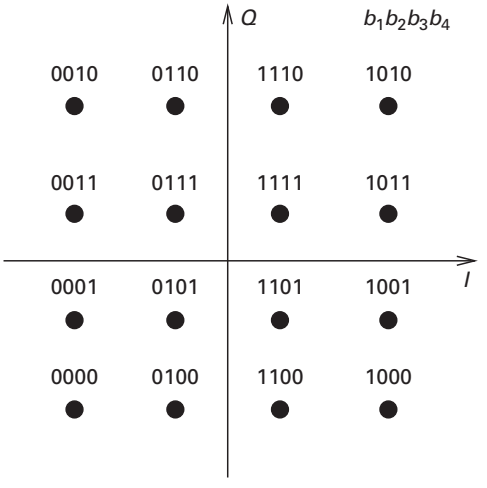


Figure 1.6 A 16-QAM constellation with Gray coding.

where K_{MOD} is a constant depending on the modulation used and the output power to be generated. An example of a 16-QAM constellation with Gray digital encoding is shown in Figure 1.6. Typical values found for the various OFDM parameters in various standards are shown in Table 1.1.

Let us discuss now how the self-interference in a given symbol is mitigated in the presence of multi-path signals. Orthogonality between the subcarriers is preserved if the symbols are periodically extended inside the guard interval time. This is symbolically represented in Figure 1.7 for the case of the reception of a symbol with three different propagation delays. In the useful symbol interval of duration T_{IFFT}, the various path delays only introduce a frequency-dependent phase shift and amplitude rescaling in

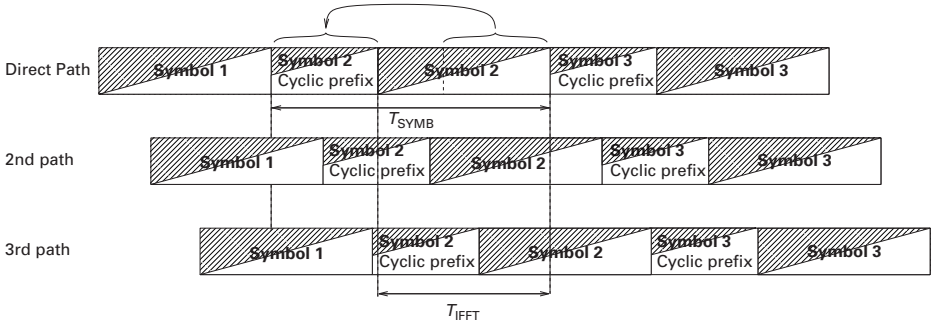


Figure 1.7 Use of a cyclic prefix to mitigate multi-path propagation.

each of the measured subcarriers which can be removed by subsequent processing at the receiver.

Having presented the fundamentals of the OFDM scheme, let us now discuss its signal characteristics. OFDM signals exhibit a very high peak-to-average power ratio (PAPR),

$$\mathrm{PAPR} = \frac{P_{\mathrm{peak}}}{P_{\mathrm{avg}}}.$$

This is due to the fact that OFDM signals are realized as the superposition of multiple subcarriers N_{SC}, and there exists a finite probability that at some times all these subcarriers can be in phase. Under such a circumstance the peak power is proportional to N_{SC}^2. Since the average power is proportional to N_{SC}, assuming all subcarriers have the same power, it results that at these fleeting times the PAPR reaches N_{SC}, which can be very large, e.g. 2048 (33 dB). Such a PAPR would be quite alarming. However, the probability of such an event is small, since it is proportional to $2^{-N_{\mathrm{SC}}}$.

Let us introduce an instantaneous PAPR(t) in a time interval $[t,\ t + \Delta t]$ defined as

$$\mathrm{PAPR}(t) = \frac{\max_{[t,t+\Delta t]}\left[|x_{\mathrm{RF}}(t)|^2\right]}{E\left[|x_{\mathrm{RF}}(t)|^2\right]},$$

with $E[\ ]$ the average operator taken over the transmit time interval of interest. It is possible to estimate the PAPR at the baseband signal level. We call $x(n)$ the complex baseband signal, which is defined as

$$x(n) = I(n) + jQ(n),$$

where the signals $I(n) = I(t_n)$ and $Q(n) = Q(t_n)$ are sampled at discrete time intervals $t_n = nt_{\mathrm{s}}$ with t_{s} the sampling interval:

$$x_{\mathrm{RF}}(t) = I(t)\cos(\omega_{\mathrm{RF}}t) - Q(t)\sin(\omega_{\mathrm{RF}}t).$$

The instantaneous PAPR at the sampling time t_n is then defined by

$$\mathrm{PAPR}(n) = \frac{|x(n)|^2}{E\left[|x(n)|^2\right]} = \frac{I^2(n) + Q^2(n)}{E_{\mathrm{avg}}} = \frac{P(n)}{P_{\mathrm{avg}}},$$

where the average $E[\ ]$ is taken over the transmitted data sequence. Note that E_{avg} is the average envelope value, $P(n)$ the discrete-time instantaneous RF power at time t_n, and P_{avg} the average RF power. Note that the instantaneous RF power $P(n)$ is simply obtained from the envelope $E(n) = |x(n)|$ using

$$P(n) = \frac{1}{2}|x(n)|^2 = \frac{1}{2}\left[I^2(n) + Q^2(n)\right] = \frac{1}{2}E^2(n).$$

Since the OFDM data can be assumed to be stochastically time-varying, the PAPR is best represented by its statistical distribution. It is common practice to use the complementary cumulative distribution function (CCDF = 1 − CDF) of the PAPR to represent its statistical distribution:

$$\text{CCDF(PAPR)} = \text{Probability}[P(n) > P_{\text{avg}} \times \text{PAPR}].$$

A simple and useful approximation for the CCDF has been proposed for QPSK OFDM [5]:

$$\text{CCDF(PAPR)} = 1 - \text{CDF} = 1 - \left[1 - \exp(-\text{PAPR})\right]^{\beta N_{\text{SC}}},$$

where the factor β is a fitting parameter used to approximate the oversampling factor. Note that $\beta = 1$ corresponds to the case of a Nyquist sampled signal. In such a Nyquist limit, by virtue of the central limit theorem, the samples $x_k(n)$ for each subcarrier can be assumed, for a large number N_{SC} of subcarriers, to be zero-mean Gaussian complex variables $N(0, \sigma^2)$ with variance σ. It results that $|x_{\text{OFDM}}(n)|$ is a Rayleigh distribution (Rayleigh[σ]) and the instantaneous power $|x_{\text{OFDM}}(n)|^2$ is exponentially distributed with mean power $\sigma^2_{\text{OFDM}} = 2\sigma^2$. In the case of oversampling, the adjacent samples $x(n)$ are correlated, and a non-unity β factor ($\beta = 2.8$) is found to provide a correction to the correlation [5]. The results obtained for various numbers of subcarriers are shown in Figure 1.8. The larger the number of subcarriers, the larger the probability of obtaining a specific PAPR.

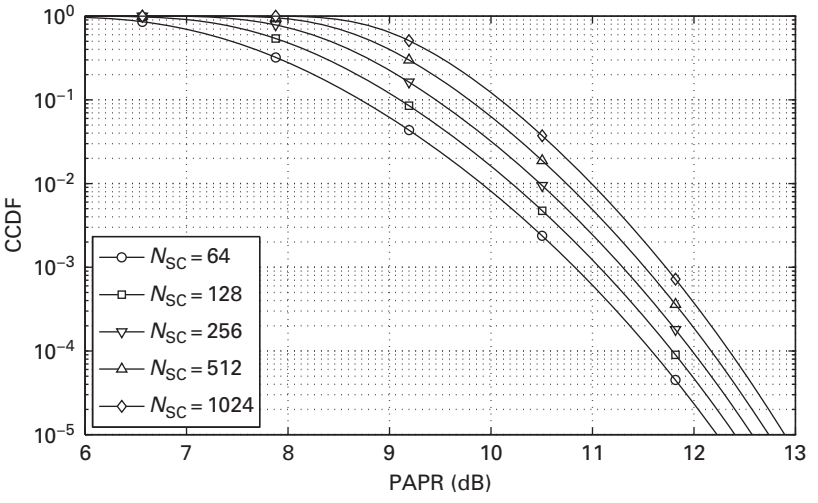


Figure 1.8 Approximate CCDF of PAPR for QPSK OFDM signals.

Note, however, that the discrete-time PAPR may underestimate the continuous-time PAPR since it is calculated using sampled data, but it nonetheless gives a useful lower-bound estimate of the analog PAPR which the amplifier must handle.

1.3 Impact of clipping on OFDM

As we have seen in the previous section, OFDM signals, just like CDMA signals, exhibit large PAPR. The question arises as to whether the peaks in the signal envelope carry critical modulation information and whether we could clip them off without significantly degrading the information transfered.

In hard clipping, the input signal $x(n) = I_{\text{in}}(n) + jQ_{\text{in}}(n)$ will have its envelope $|x(n)| = \sqrt{I_{\text{in}}^2(n) + Q_{\text{in}}^2(n)}$ rescaled to E_{clip} if the envelope exceeds the clipping threshold E_{clip}. However, this envelope clipping is performed in such a way as to maintain the phase $\angle(I_{\text{in}} + jQ_{\text{in}})$ information of the input signal. The output signal $y(n) = I_{\text{out}}(n) + jQ_{\text{out}}(n)$ will therefore be given by

$$I_{\text{out}}(n) = \text{Re}[x(n)] = \frac{E_{\text{clip}}}{|x(n)|} I_{\text{in}}(n),$$
$$Q_{\text{out}}(n) = \text{Im}[x(n)] = \frac{E_{\text{clip}}}{|x(n)|} Q_{\text{in}}(n).$$

Note that the output phase satisfies $\angle(I_{\text{out}} + jQ_{\text{out}}) = \angle(I_{\text{in}} + jQ_{\text{in}})$. Figure 1.9(a) shows a processing block that conceptually implements this hard-clipping function using an antipeak generator. The output signal $y(n) = x(n) + c(n)$ is the superposition of the incident signal x_{in} and a corrective clipping signal $c(n)$ generated by the antipeak generator:

$$c(n) = I_{\text{C}} + jQ_{\text{C}} = (I_{\text{out}} - I_{\text{in}}) + j(Q_{\text{out}} - Q_{\text{in}}).$$

The clipping ratio γ is defined as

$$\gamma = \frac{E_{\text{clip}}}{\sqrt{E[|x(n)|^2]}} = \frac{E_{\text{clip}}}{E_{\text{avg}}} = \sqrt{\frac{P_{\text{clip}}}{P_{\text{avg}}}},$$

where for OFDM signals the average power is $P_{\text{avg}} = \sigma_{\text{OFDM}}^2$.

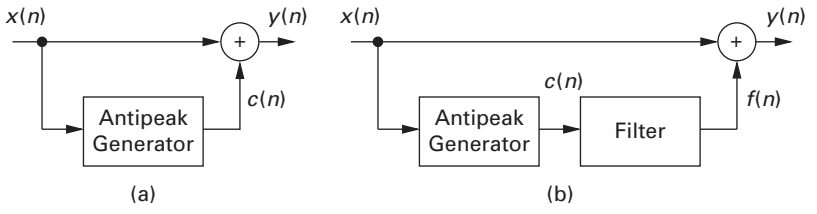


Figure 1.9 System diagrams for implementing hard clipping (a) and soft clipping (b).

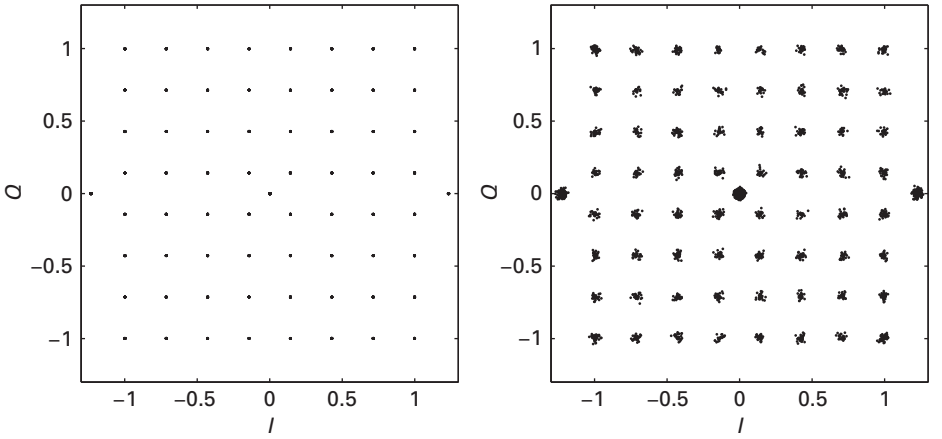


Figure 1.10 Constellation for the various OFDM subcarriers before (left) and after (right) hard clipping with a clipping level of 7 dB.

Hard clipping distorts the OFDM signals and generates both out-of-band spectral regrowth and in-band distortion. Let us first consider the in-band distortion. The distortion of the constellation in 64-QAM OFDM is shown in Figure 1.10. Clipping is seen to have introduced noise. Although the corrective signal $c(n)$ is clearly correlated to the input signal x_{in}, it has been demonstrated [6] [7] that for OFDM signals with a large number of subcarriers the clipped signal can be well represented by the superposition of an attenuated version of the input signal $\alpha x(n)$ and an uncorrelated additive noise $d(n)$:

$$y(n) = \alpha x(n) + d(n),$$

with α given by

$$\alpha = 1 - \exp(-\gamma^2) + \frac{\sqrt{\pi}}{2}\gamma \operatorname{erfc}[\gamma].$$

Figure 1.11 shows the variation of the attenuation constant α with the clipping ratio γ. For clipping ratios above 7 dB, the attenuation constant α is close to unity and can safely be neglected.

The uncorrelated additive noise $d(n)$ has been demonstrated [7] to have a variance given by

$$\sigma_{\text{d}}^2 = \sigma_{\text{OFDM}}^2 \left(1 - e^{-\gamma^2} - \alpha^2\right).$$

The bit error probability (BEP) for M-QAM OFDM under clipping has been derived [7] to be given by

$$\text{BEP}_{\text{OFDM}} = \frac{2}{\log_2(M)} \left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left[\frac{3 \times \text{SNDR}}{2(M-1)}\right],$$

with $\log_2(M)$ the number of bits in each Mary-QAM subcarrier. In the expression for the BEP, SNDR is the signal-to-noise-plus-distortion ratio:

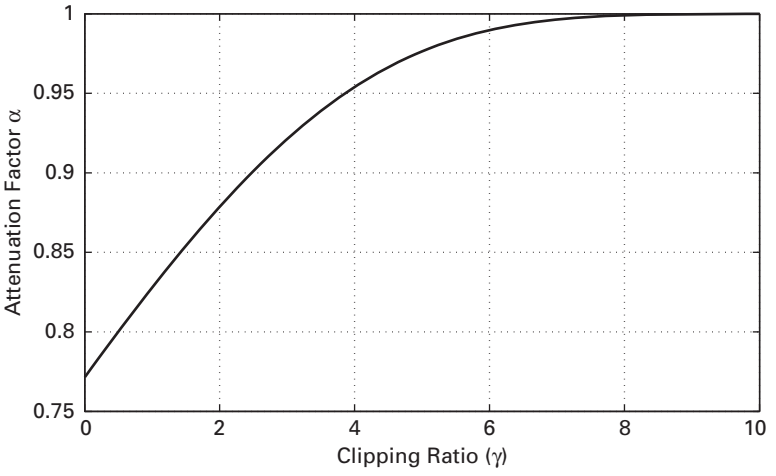


Figure 1.11 Variation of the attenuation constant α versus the clipping ratio γ.

$$\mathrm{SNDR} = \frac{\alpha^2 \sigma^2_{\mathrm{OFDM}}}{\sigma^2_{\mathrm{d}} + \sigma^2_0},$$

where σ_0 is the variance associated with the thermal noise floor:

$$k_{\mathrm{B}} T_0 = N_0 = \frac{1}{2}\sigma_0^2 T_{\mathrm{SYMB}},$$

with k_{B} the Boltzmann constant, T_0 the room temperature in degrees Kelvin, $N_0/2$ the additive white-noise spectral density, and T_{SYMB} the symbol duration. For an Mary-QAM OFDM signal, the ratio E_{b}/N_0 of the average energy per bit to N_0 is related to the average power σ^2_{OFDM} of the OFDM signal by

$$\frac{E_{\mathrm{b}}}{N_0} = \frac{\alpha^2}{\log_2(M)} \frac{\sigma^2_{\mathrm{OFDM}}}{\sigma^2_0}.$$

The BEP of a clipped OFDM signal is plotted for various clipping levels from 4 to 10 dB in Figure 1.12. The OFDM signal consists of $N_{\mathrm{SC}} = 2048$ subcarriers using 64-QAM modulation ($M = 64$). It is seen that for a clipping threshold equal to or larger than 7 dB (above the average power), an acceptable degradation in BEP is obtained.

Note that various techniques for PAPR reduction without BEP degradation at the cost of bandwidth efficiency have been proposed. Although these investigations fall outside the scope of this review, this points to the fact that the characteristics of the OFDM signals will vary from implementation to implementation even within the same standard.

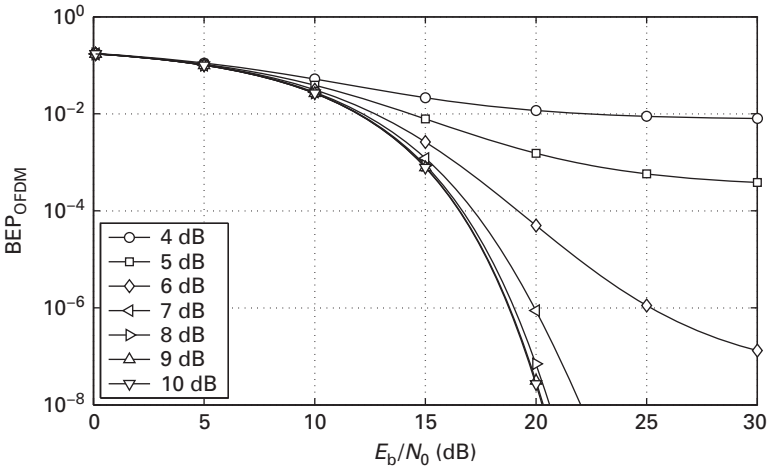


Figure 1.12 Bit error probability for various clipping levels varying from 4 to 10 dB.

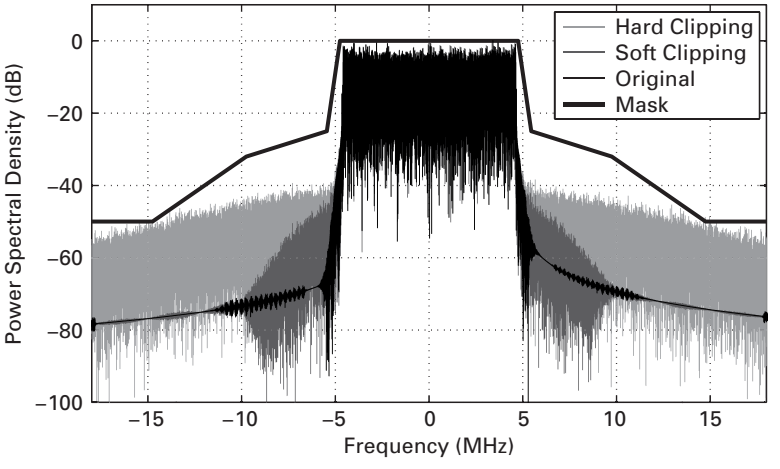


Figure 1.13 Comparison of the power spectral densities of the original, hard-clipped, and soft-clipped signals.

1.4 Spectral regrowth and clipping

As is to be expected, hard clipping will increase the modulation bandwidth and introduce spectral regrowth outside the allocated bandwidth. This is illustrated in Figure 1.13, where the power spectral densities of the original and hard-clipped signals are compared. Soft clipping can then be used to reduce the spectral regrowth. In soft clipping, a filter is used after the antipeak generator to remove the sharp variations in the correction signal $c(n)$, resulting in the implementation shown in Figure 1.9(b). The resulting power spectral density (PSD) of the soft-clipped signal is compared in Figure 1.13 with the PSDs of the original and hard-clipped signals for a clipping

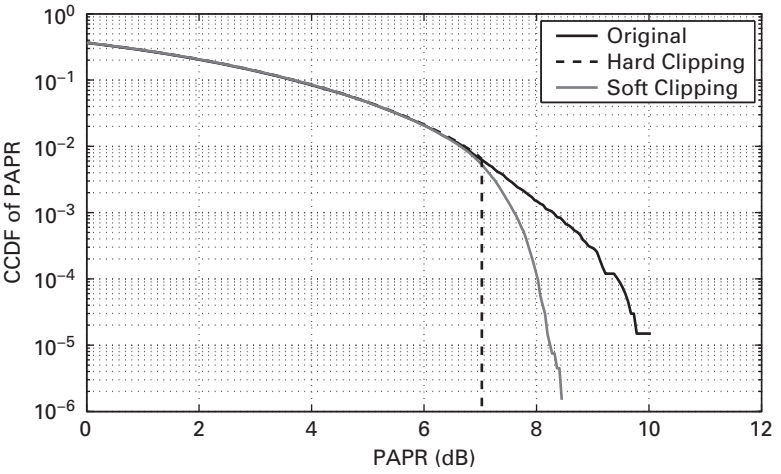


Figure 1.14 Comparison of the CCDFs of the original, hard-clipped, and soft-clipped WiMax signals (10 MHz).

ratio of 7 dB. It is observed that the soft-clipping PSD easily passes the requirement of one of the proposed spectral masks for the IEEE 806.16e-2005 WiMAX standard. The associated CCDFs are also compared, in Figure 1.14.

1.5 Metrics

To compare the linearity performance of different hardware implementations, or to compare the fidelity of a model in fitting or predicting measured data, several metrics are commonly used.

For comparing two signals $x_1(n)$ and $x_2(n)$ available in the discrete time domain, the root mean square (RMS) is often used. It is defined as

$$\mathrm{RMS} = \sqrt{\frac{\sum_n |x_2(n) - x_1(n)|^2}{\sum_n |x_1(n)|^2}}.$$

To measure the spectral regrowth generated in adjacent bands ($\omega_{\mathrm{adjacent}}$) by the signal in the channel band ($\omega_{\mathrm{channel}}$), the adjacent-channel power ratio (ACPR) is used. It is defined from the power spectral density PSD as

$$\mathrm{ACPR} = 10\log_{10}\left\{\frac{\int_{\omega_{\mathrm{adjacent}}} \mathrm{PSD}(\omega)d\omega}{\int_{\omega_{\mathrm{channel}}} \mathrm{PSD}(\omega)d\omega}\right\}$$

For comparing, in symbol-space, two similar OFDM signals $x_{\mathrm{OFDM},1}$ and $x_{\mathrm{OFDM},2}$ with N_{SC} subcarriers k, the following error vector magnitude (EVM) is used:

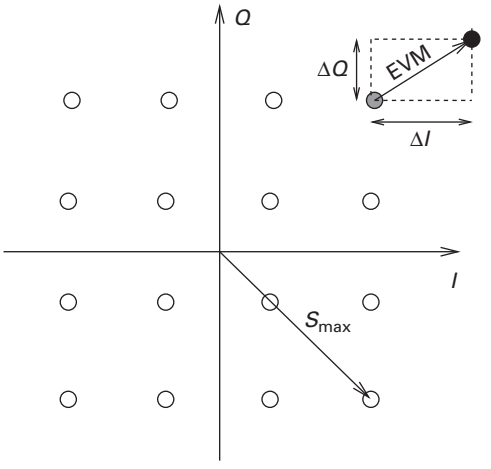


Figure 1.15 Calculation of EVM error for a given 16-QAM subcarrier k.

$$\mathrm{EVM} = \frac{1}{S_{\max}} \left(\frac{1}{N_{\mathrm{SC}}} \sum_{k=1}^{N_{\mathrm{SC}}} \Delta I_k^2 + \Delta Q_k^2 \right)^{\frac{1}{2}} = \frac{\sigma_{\mathrm{d}}}{S_{\max}},$$

with

$$\Delta I_k = I_{2,k} - I_{1,k},$$
$$\Delta Q_k = Q_{2,k} - Q_{1,k}.$$

The square root of the component $\Delta I_k^2 + \Delta Q_k^2$ of the EVM error is graphically represented in Figure 1.15 for a single 16-QAM subcarrier. Other figures of merit will also be introduced in the subsequent chapters as needed for our analysis.

1.6 Multisine

Multisine is an expression used to indicate a superposition of tones, of the following form [8]:

$$x_{\mathrm{ms}}(t) = \sum_{n=0}^{N_{\max}} A_n \cos(\omega_0 t + n\,\Delta\omega + \phi_n),$$

for single-sided (usually baseband) signals, or

$$x_{\mathrm{ms}}(t) = \sum_{n=-\mathrm{SSB}}^{\mathrm{SSB}} A_n \cos(\omega_0 t + n\,\Delta\omega + \phi_n),$$

for double-sided RF signals with $2 \times \mathrm{SSB} + 1$. Usually the same amplitude A_n is used and only the phases ϕ_n are varying for the tones n.

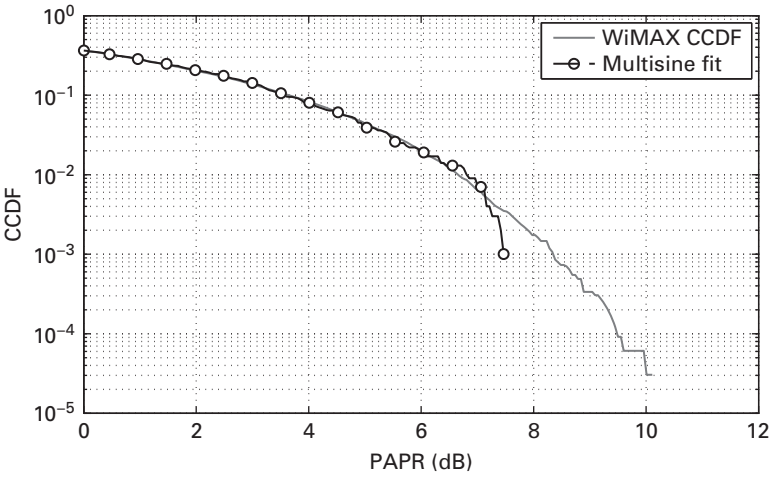


Figure 1.16 Approximation of a 1024-carrier WiMAX CCDF using an optimized 64-tone multisine.

The selection of the phases ϕ_n is of critical importance for shaping the multisine signal [8] [9]. Using phases that are constant or linearly varying or randomly distributed leads to pulse-like signals with large PAPR. To approximate communication signals using multisines, the phase ϕ_m can be optimized such that the desired CCDF is approached [10] [11]. An example of a CCDF approximation is shown in Figure 1.16 for 64 tones. The more tones used by the multisine, the closer the CCDF can be approximated at high PAPR. Nevertheless, a multisine with a reduced number of tones might provide an acceptable substitute for soft-clipped OFDM signals. Such multisines find applications in the characterization and modeling of nonlinear devices with memory effects [12].

References

[1] B. Razavi, *RF Microelectronics*, Prentice Hall, 1998.

[2] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX, Understanding Broadband Wireless Networking*, Prentice Hall, 2008.

[3] F. Khan, *LTE for 4G Mobile Broadband, Air Interface Technologies and Performance*, Cambridge University Press, 2009.

[4] S. H. A. Shah, M. Iqbal, and T. Hussain, "Comparison between WiMAX and 3GPP LTE," Master of Science Thesis, Blekinge Institute of Technology, 2009.

[5] R. van Nee and A. de Wild, "Reducing the peak to average power ratio in OFDM," in *48th IEEE Vehicular Technology Conference*, pp. 2072–2076, 1998.

[6] H. E. Rowe, "Memoryless nonlinearities with Gaussian inputs: elementary results," *Bell Systems Technical Journal*, Vol. 1, No. 7, pp. 1519–1525, Sept. 1982.

[7] D. Dardari, V. Tralli, and A. Vaccari, "A theoretical characterization of non-linear distortion effects in OFDM systems," *IEEE Transactions on Communications*, Vol. 48, No. 10, pp. 1755–1764, Oct. 2000.

[8] K. A. Remley, "Multi-sine excitation for ACPR measurements," in *IEEE MTT-S International Microwave Symposium*, pp. 2141–2144, 2003.

[9] J. C. Pedro and N. B. de Carvalho, "On the use of multitone techniques for assessing RF components' intermodulation distortion," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 12, pp. 2393–2402, Dec. 1999.

[10] M. B. Steer, "Accurate multisine representation of digital communication signals for characterization of nonlinear circuits," in *2006 Radio and Wireless Symposium*, pp. 527–530, 2006.

[11] S. Jiangtao, J. Benedikt, and P. J. Tasker, "A new method for the design of multi-sine excitations for the assessment of non-linear devices," in *2008 MTT-S International Microwave Symposium Digest*, pp. 265–268, 2008.

[12] D. Schreurs and K. A. Remley, "Use of multisine signals for efficient behavioral modeling of RF circuits with short-memory effects," in *61st ARFTG Conference Digest*, pp. 65–72, 2003.

2 Large-signal vector measurement techniques with NVNAs[1]

In this chapter we shall first review the various measurement techniques which are traditionally used for the characterization of RF circuits. Next we will introduce several nonlinear vector network analyzers (NVNAs) that have been developed to characterize the nonlinear response of circuits at radio frequencies (RF). Having described the operating principle of the various NVNAs, we will then focus in the rest of this chapter on the sampler-based NVNA. In particular, its operating principle, calibration, and extension to broadband modulation and pulsed RF signals will be discussed in detail. Overall, this chapter will provide some insights into nonlinear measurement techniques, which should complement the remaining chapters concerned with platform-independent modeling, design, and linearization techniques.

2.1 Measurement of RF signals

Various kinds of equipment are used to acquire and analyze RF signals: power meters, spectrum analyzers, oscilloscopes, vector signal analyzers. Power meters are used to accurately measure the RF power. Usually they are used to characterize steady-state RF signals but some power meters have a wide input bandwidth and can provide a measurement of the instantaneous signal power for modulated RF signals within a prescribed bandwidth. The accuracy of power meters is usually traceable to world standards and such equipment can therefore be used for power calibration.

Spectrum analyzers permit the measurement of the power spectral density of RF signals versus frequency. Typically the data are acquired during a finite-duration time window and Fourier transformed to the frequency domain for a wide frequency range. The frequency resolution f_{RES} improves (decreases) with increased measurement duration $t_{\mathrm{M}} = 1/f_{\mathrm{RES}}$ at the cost of slower frequency sweep. No phase information is normally provided by spectrum analyzers.

Oscilloscopes permit visualization of the data in the time domain. RF oscilloscopes need to have a sampling rate of at least twice (Nyquist rate) the signal frequency in order to acquire the fundamental. For acquiring the 10th harmonic of a 1-GHz signal

[1] Research collaboration with Seok Joo Doo, Young Seo Ko, Fabien De Groote, Jean-Pierre Teyssier, and Chieh Kai Yang is gratefully acknowledged.

we therefore need a sampling rate of 20 GHz. Oscilloscope measurements are usually affected by the jitter of the sampling clock, but techniques are available to compensate for this jitter [1]. Scopes have also a finite frequency response and typically feature a resolution of 8 bits (48 dB dynamic range).

Vector signal analyzers (VSAs) are used to analyze the modulation of RF signals. These instruments act as calibrated receivers that demodulate the RF signal received. Typically they can display the constellation diagram and calculate the error vector magnitude (EVM) between intended and received signals. Only the fundamental frequency is characterized by such systems. A rigorous calibration of a VSA requires a pre-calibrated modulated RF source.

The various types of test equipment described above are intended to measure CW and modulated RF signals and can be used in turn to characterize circuits, as we shall discuss below. First let us consider the case of linear circuits, or more accurately the linear characterization of circuits, since no circuit is ever purely linear. A dedicated piece of equipment, the network analyzer, is available for this purpose. In a network analyzer, commonly two-port or four-port circuits are excited by an incident RF wave. The power level must be selected by the user so that the device response is linear while remaining above the noise floor both of the device and of the measurement system. Network analyzers have typically a dedicated RF synthesizer that is capable of rapidly sweeping the frequency of the RF excitation. Network analyzers do not per se acquire the incident and reflected or transmitted waves but evaluate their ratios:

$$S_{ij}(\omega) = \left.\frac{b_i(\omega)}{a_j(\omega)}\right|_{a_k=0 \text{ with } k\neq j},$$

where $b_i(\omega)$ are the reflected or transmitted complex waves and $a_j(\omega)$ the incident complex waves on the DUT at the radial frequency ω. For linear circuits sufficient information is provided by the knowledge of the amplitude and phase of these S-parameters. Given the characteristic impedance, S-parameters can indeed be converted into other circuit parameters such as z, y, and $ABCD$ parameters. For nonlinear circuits, care must be taken to use a low enough incident power to measure small-signal parameters.

To characterize the nonlinear response of circuits testbeds have been developed to perform nonlinear measurements, such as power gain compression, power efficiency, AM–AM and AM–PM, intermodulation distortion (IMD), load-pull measurements, and EVM measurements.

Figure 2.1 shows examples of these various testbeds. A power meter, a spectrum analyzer, and a vector signal analyzer are required in turn to measure the signal sent or transmitted to the device/circuits measured. A network analyzer is required in order to pre-characterize the automatic tuner or measure the source and load impedances in manual load-pull testbeds. Note that some network analyzers have been configured with power sweep capability and offset frequency tuning for measuring the amplitude of higher harmonics.

To the extent that the figures of merit acquired are the very metrics used to measure the success of the circuits tested, these testbeds have proven invaluable. However,

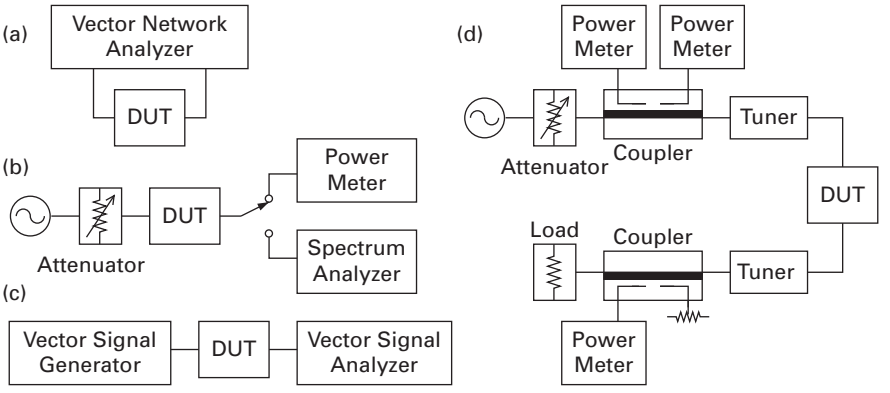


Figure 2.1 Various testbeds used for device and circuit characterization: (a) network analyzer used for S-parameters parameters, (b) power meter and spectrum analyzer for power-swept measurement of gain compression and harmonics and spectral regrowth, (c) vector signal analyzer for constellation acquisition and EVM, and (d) load-pull system for PA development.

comparing measurements acquired from different testbeds may pose a serious challenge if one attempts to reconcile them for accurate device modeling, since different calibrations are typically used. In addition, the harmonic termination impedance might be different. Even different low-frequency impedance termination will affect the device and circuit RF response. Furthermore, designers are strongly interested in directly visualizing the voltage and current waveforms at the reference planes of the device under test (DUT), and these traditional methods alone do not provide the required vectorial (amplitude and phase) multi-harmonic information needed to reconstruct those signals.

In this book on nonlinear circuits we shall be concerned with measurement systems capable of performing vector multi-harmonic measurements so as to visualize in time, frequency or envelope domains the full waveforms at the DUT. As we shall see, a family of instruments is already available, and several novel exciting low-cost solutions are emerging, which will soon make such nonlinear vector measurements ubiquitous.

2.2 Principle of operation of vector large-signal measurements

Several test systems have been conceived to measure the nonlinear vector response of RF devices and circuits. These test systems can be separated into three categories: sampling-based, mixer-based, and sub-sampling-based systems. In this book we shall refer to them all as nonlinear vector network analyzers.

The sampling approach simply relies on high-giga-sample oscilloscopes. Such testbeds have been developed in the industry and universities by experts [2] in RF nonlinear measurements, and are being commercialized by companies such as Mesuro. The prototype for such a system is shown in Figure 2.2. The RF incident and reflected waves of the DUT are separated by a directional coupler (reflectometer) and directly measured by a four-channel sampling oscilloscope. A high-bandwidth

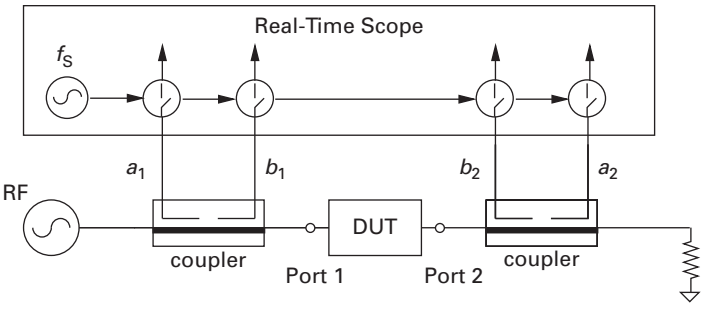


Figure 2.2 NVNA realized with a high-bandwidth high-speed sampling scope.

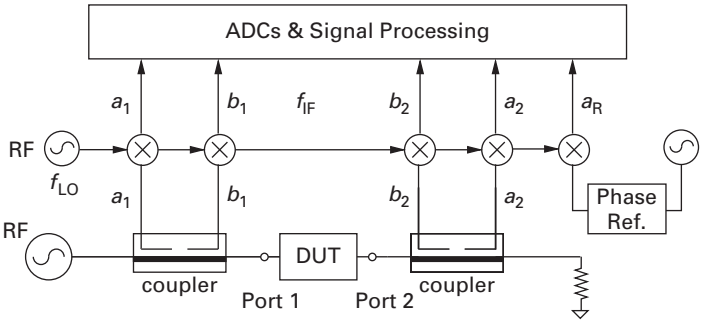


Figure 2.3 Architecture for a mixer-based NVNA realized with a five-channel VNA.

scope (e.g. 67 GHz) with a large sampling rate compared with the fundamental RF frequency (typically a few GHz) measured is used, in order to achieve a sufficiently high oversampling to permit the faithful acquisition of higher harmonics. The jitter of the sampling clock will affect the measurement, but techniques to remove its contribution are available [1]. However, jitter removal in such testbeds requires a fifth channel for the acquisition of a pilot sinusoidal tone. Relative calibration and absolute power calibration should be sufficient to characterize the coupler and bias tee loss and account for the frequency response of the scope. No phase calibration is conceptually required when the four channels are fully synchronized. This is certainly verified if we characterize relatively low RF harmonic frequencies compared with the scope bandwidth. The dynamic range of the system is limited by that of the scope. With an 8-bit vertical resolution about 48 dBc is obtained. In principle such a system is not limited to periodic signals, but, when measuring periodic signals, averaging can be used to further increase the dynamic range at the cost of longer measurements and increased memory storage.

Mixer-based vector nonlinear vector network analyzers (NVNAs) are usually built on a five-channel network analyzer standard platform (Figure 2.3). Several implementations have already been developed (by Barataud *et al*. [3] and Phaser-Quattro [4]) in university labs. Several solutions have been marketed (PNAPlus and ZVxPlus) by NMDG [5] as extensions of commercially available network analyzers, and a new

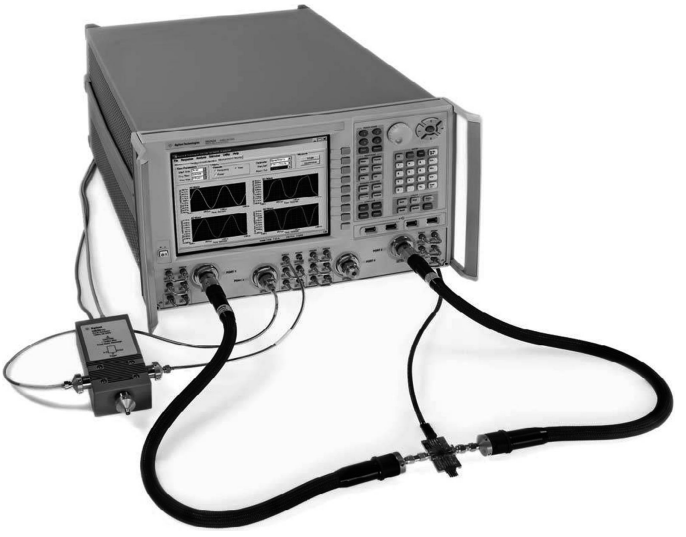


Figure 2.4 PNA, a mixer-based NVNA, commercialized by Agilent (reproduced with permission from Agilent Technologies, Inc.).

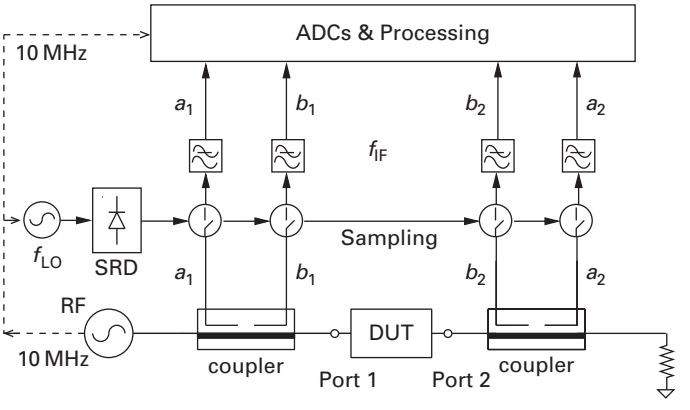


Figure 2.5 Sampler-based NVNA architecture.

system based on the PNA-X architecture is now being commercialized by Agilent [6]. The mixer-based NVNA provides a high-performance solution. Figure 2.4 shows a prototype for such a system. The RF incident and reflected waves a_1, b_1, a_2, and b_2 of the DUT which are separated by the directional coupler (reflectometer) are down converted by a mixer using the local frequency $f_{LO} = f_{RF} + f_{IF}$. The resulting IF signals are then acquired by a sampler (Figure 2.5). A narrow bandwidth can be selected to reduce the noise at the price of longer acquisition time. Each harmonic (or sideband when the signal is modulated) is measured in separate individual measurements. The RF signals measured must therefore be periodic and stable. Correlation between these measurements at different frequencies is obtained by measuring on the fifth channel a multi-harmonic signal generated by the synchronizer generator. The phase relationship of the harmonic

Figure 2.6 LSNA, a sampling-based NVNA.

tones generated by the synchronizer does not need to be known as long as it is fixed and stable. The calibration of the mixer-based NVNA relies on a relative calibration and absolute power and phase calibration of the same type as those we shall describe in detail for the sub-sampling NVNA. The advantage of the mixer-based NVNA is that it offers a higher dynamic range (80–90 dB) than that of the sampling and sub-sampling NVNA. The resolution (100 kHz for the Phaser-Quattro) is maintained across the entire bandwidth, and the lower frequency limit demonstrated is lower (300 kHz in the Quattro versus 600 MHz in the Agilent Large Signal Network Analyzer). The mixer-based approach also benefits from the fast frequency-sweeping capability of modern network analyzers, although, for the acquisition of modulated RF signals with a large number of sidebands, the measurement will require a commensurately longer acquisition time since each tone is acquired in a separate measurement.

The sub-sampling NVNA shown in Figure 2.6 relies on the sampling of the RF signals at a sampling frequency much lower (typically around 20 MHz) than the RF signal measured (typically from 0.6 to 50 GHz).

Whereas the sampling-based and mixer-based NVNAs down convert the signal in the time and frequency domains, respectively, the sub-sampling NVNA works in both time and frequency domains. Indeed, for this approach to work the RF signals measured (and their modulation) must be periodic as in the mixer-based NVNA in order for the sampling to perform an effective down conversion of the signals. The sampling frequency must be selected such that the down-converted tones do not overlap in the frequency domain, as will be discussed in the next section. The Large Signal Network Analyzer (LSNA) [7] is an example of a sub-sampling NVNA. Its hardware is based on the older Microwave Transition Analyzer (MTA) sampling boards. The reader is

Figure 2.7 The SWAP, a new sampling-based NVNA with improved performance. (Reproduced with permission from VTD sas.)

referred to reference [8] for a historical review of the development of the large-signal network analysis. Although the MTA and the LSNA have now both been discontinued, a sampler-based NVNA, the SWAP (see Figure 2.7) is now commercialized by VTD [9]. The SWAP provides a low-cost replacement solution for the LSNA while featuring an improved performance in terms of speed of acquisition, modulation bandwidth, and pulsed-RF measurement capabilities.

Clearly our brief overview of various NVNAs available indicates that a diversity of solutions is available for users interested in large-signal measurements.

2.3 Sampler-based principle of operation

Having introduced the major players, we shall now focus on the sampler-based NVNAs (MTA, LSNA, SWAP) for the rest of this chapter. Note also that a wide variety of measurement configurations can be explored with the sampler-based NVNA because the user fully determines the RF and baseband stimuli as long as their modulations are periodic.

Note that our discussion is intended for instrument users, not test-instrument developers. The development of instruments entails many considerations that users may ignore without suffering any practical limitations. However, to use the sampler-based NVNA with any reasonable success, a basic understanding of the fundamental principles of the instrument is required. We shall therefore start by reviewing its architecture and operating principles.

Typically the sampler-based NVNA, which is intended for measuring two-port RF devices, features four independent channels each equipped with a sampler. The sampling clock driving the four samplers is generated by an extremely precise and stable source (a so-called FracN source) which is programmable (typically 10–25 MHz).

The FracN source is phase-locked to a stable (oven-stabilized) 10-MHz reference signal (typically provided by the highest-frequency RF source). The sampling impulses are realized with a step recovery diode producing a short pulse of duration a few picoseconds. The four sampled RF signals which have been effectively down converted to baseband via sub-sampling are then filtered by a 10-MHz filter (50 MHz for the SWAP) and acquired by the PC by four analog-to-digital converters (ADCs). The latter ADCs are also phase-locked to the 10-MHz reference clock signal for increased signal stability.

To illustrate the concept of sub-sampling down conversion, let us consider the following example. Let us assume we are to measure a signal with fundamental frequency $f_{\mathrm{RF}} = 1\,\mathrm{GHz}$ and $N_{\mathrm{H}} = 10$ harmonics:

$$f_{\mathrm{RF},n} = nf_{\mathrm{RF}} \qquad \text{with} - N_{\mathrm{H}} \leq n \leq N_{\mathrm{H}}.$$

The RF signal might be modulated later on, so we need to decide on a resolution frequency for these modulated measurements. Assuming that the modulation frequency can be as low as 1 kHz, let us select a resolution frequency of $f_{\mathrm{RES}} = 100\,\mathrm{Hz}$ in this example.[2] It results that the RF signal frequency is $R_{\mathrm{RF}} = 10^7$ times the resolution frequency:

$$f_{\mathrm{RF}} = R_{\mathrm{RF}} f_{\mathrm{RES}},$$

assuming for simplicity that f_{RES} is an integral divider for f_{RF}. Note that, to achieve such a resolution frequency, we need a measurement time of $1/f_{\mathrm{REF}} = 1/(100\,\mathrm{Hz}) = 10\,\mathrm{ms}$. If we were using a sampling scope capable of both sampling and storing the RF data at high enough speed, we would need, according to the Nyquist criteria, to acquire $2 \times R_{\mathrm{RF}} \times N_{\mathrm{H}} = 2 \times 10^7 \times 10 = 2 \times 10^8$ sampled data during this 10-ms measurement time. To obtain the RF spectrum with the required acquired data we would then need to perform a fast Fourier transform (FFT) with 2×10^8 data. This would require tremendous computational resources, even assuming that we can practically store so much data fast enough. With sub-sampling, fewer data need be acquired to perform the same task. First we need to find the sampling frequency f_{LO} which down converts the RF signal f_{RF} and its harmonics nf_{RF} to baseband tones such that their down-converted frequencies in the sampling frequency interval,

$$0 \leq f_n = nf_{\mathrm{RF}} - S_n f_{\mathrm{LO}} \leq f_{\mathrm{LO}},$$

with n an integer, do not overlap after down conversion. Using a simple search algorithm, one can easily verify that the smallest sampling frequency possible for our selected example is given by

$$f_{\mathrm{LO}} = R_{\mathrm{LO}} f_{\mathrm{RES}} = f_{\mathrm{BW}},$$

[2] The resolution frequency is the inverse of the measurement time. Note that in practice, in the sampler-based NVNA, the minimum resolution frequency is set by the ADC sampling frequency divided by the maximum number of sampled points acquired during the measurement time.

with $R_{LO} = 21$. One easily verifies that in the general case this corresponds to the minimum positive frequency window required to collect the down-converted tones including positive and negative (Nyquist image) frequencies:

$$f_{BW} = (2N_H + 1) f_{RES}.$$

Let us analyze this process in more detail. We assume that the unmodulated RF signal is given by the following Fourier series:

$$x(t) = \sum_{k=0}^{N} C_k e^{jk\omega_0 t}.$$

Let us assume that we can directly sample the RF signal using the following sampling-function signal:

$$S(t) = \sum_{n=-\infty}^{\infty} \delta(t - nt_S).$$

We now calculate the pth Fourier coefficient F_p of the sampled data $x(t) \times S(t)$:

$$\begin{aligned}
F_p &= \int_{0^+}^{t_R^+} [x(t) \times S(t)] e^{jp\omega_{RES}t}\, dt \\
&= \int_{0^+}^{t_R^+} \left[\sum_{k=0}^{N} C_k e^{jk\omega_0 t} \right] \times \left[\sum_{n=-\infty}^{\infty} \delta(t - nt_S) \right] e^{jp\omega_{RES}t} dt \\
&= \sum_{k=0}^{N} C_k \sum_{n=-\infty}^{\infty} \int_{0^+}^{t_R^+} \exp\big[j(k\omega_0 + p\omega_{RES})t\big] \delta(t - nt_S) dt \\
&= \sum_{k=0}^{N} C_k \sum_{n=1}^{t_R/t_S} \exp\big[j(k\omega_0 + p\omega_{RES})nt_S\big] = \sum_{k=0}^{N} C_k G(k, p),
\end{aligned}$$

where we define G as

$$G(k, p) = \sum_{n=1}^{t_R/t_S} \exp[jn2\pi M(k, p)],$$

with the function M given by

$$M(k, p) = \left(\frac{k\omega_0 + p\omega_{RES}}{\omega_S} \right).$$

When M is an integer, the phasors are all equal to unity and the G function adds up to t_R/t_S. For non-integer values of M the phasors are equally distributed over a circle of unit radius and cancel each other out. It results that the Fourier coefficients obtained are weighted by f_{RES}/f_{LO}:

$$F_{p(k)} = \frac{f_{RES}}{f_{LO}} C_k,$$

with the function $p(k)$ given by

$$p(k) = \frac{k\omega_0 - \text{floor}(k\omega_0/\omega_S) f_{LO}}{\omega_{RES}} = kR_0 - \text{floor}\left(k\frac{R_0}{R_{LO}}\right) R_{LO},$$

where we defined $f_0 = R_0 f_{RES}$ and $f_{LO} = R_{LO} f_{RES}$.

Let us assume now that each RF harmonic is modulated by a modulation frequency f_m with N_{SSB} single-sideband (SSB) tones. The resulting RF tones are then given by

$$f_{RF,n,p} = nf_{RF} + pf_m \qquad \text{with } -N_H \le n \le N_H \text{ and } -N_{SSB} \le p \le N_{SSB}.$$

We also assume for simplicity that f_{RES} is an integral divider of f_m:

$$f_m = R_m f_{RES}.$$

The minimum frequency window required to collect these down-converted tones is

$$f_{BW} = (2N_H + 1)(2N_{SSB} + 1) f_{RES}.$$

The sampling f_{LO} frequency will therefore be larger than, or at best equal to, f_{BW}. Consider for example the case in which SSB = 20. Let us assume the resolution frequency is $f_{RES} = 100$ Hz. The smallest sampling frequency leading to non-overlapping tones after down conversion is

$$f_{LO} = R_{LO} f_{RES} = 103\,700 \text{ Hz}.$$

In theory it is therefore sufficient to sample the RF signal with the sampling frequency $f_{LO} = 103\,700$ Hz during the time interval of $1/f_{RES} = 10$ ms and acquire $R_{LO} = 1037$ samples to fully characterize this modulated RF signal. The $(2N_H + 1)(2N_{SSB} + 1) f_{RES}$ Fourier coefficients of this modulated RF signal are then obtained by performing an FFT with $R_{LO} = 1037$ sampled data. This is evidently much more efficient than calculating the FFT for $N_H f_0/f_{RES} = 10^8$ sampled data. This reduction in required resources originates from the fact that the signal characterized was assumed to be bandlimited around each harmonic.

No filter is needed for direct digital sub-sampling in the ideal case in which both the number of harmonics (N_H) and the number of single sidebands (N_{SSB}) are finite such that both the RF signal and its modulation are bandlimited. In practice direct digital sampling of the RF signal is not always possible and an analog sub-sampling circuit as in Figure 2.5 is used to sample the RF signal. The zero-output signal resulting between the sampler pulses then induces the generation of Nyquist images much like in digital-to-analog converters or in interpolation by zero-padding for oversampling. To eliminate the contribution of these higher Nyquist images an IF lowpass filter (10 MHz in the LSNA) is used to prevent any aliasing.

The implementation of sub-sampling in a real system is further constrained by a number of additional practical reasons such as the frequency range of operation of the harmonic phase reference and the LO source. However, the simplified analysis provided above has the advantage of introducing the overall sub-sampling down-conversion principle as well as defining the practical concept of resolution frequency. A more realistic analysis of the down-conversion process implemented in the sampler-based NVNA

together with practical examples will be given in Sections 2.7 and 2.9 when we discuss broadband and multiple recording measurements.

For reliable and reproducible measurements of modulated RF data with the sampler-based NVNA it is necessary in practice to use a modulation frequency f_{m} defined with an accuracy Δf_{m} of at least a tenth of the resolution frequency f_{RES} (the inverse of the measurement time). One can also deduce from this criterion that the periodicity requirement on the modulated RF signal can be implemented by selecting the RF frequency and the modulation frequency independently as long as they have a common frequency divider that is smaller than the resolution frequency. However, no detailed analysis of the resulting residual phase error in such a case has been reported.

2.4 Relative and absolute power and harmonic phase calibrations

The full characterization of the modulated multi-harmonic signals to be measured by the NVNA includes a baseband component (DC included) and an RF component (fundamental plus harmonics) including the sideband tones. In this section we will focus on the RF calibration part of the LSNA.

The initial theory on the RF calibration of the LSNA was reported by Jan Verspecht in his Ph.D. dissertation [11]. A simplified and alternative presentation that will be sufficient for our purposes is given here.

It is found necessary to differentiate between two types of calibration (Figure 2.8) for (1) connectorized DUT and (2) on-wafer measurements. Indeed, for on-wafer measurements, power meters or phase-reference generators calibrated at the probe contacts are not available.

2.4.1 Calibration for connectorized devices

We shall start with the calibration of the LSNA for connectorized devices. The measurement of the incident and reflected waves $a_1(t)$, $b_1(t)$, $a_2(t)$, and $b_2(t)$ at the reference planes for all the various tones $n\omega_0$ is targeted:

$$a_i(t) = \sum_{n=-N}^{N} a_i(n\omega_0)\exp(jn\omega_0 t),$$

$$b_i(t) = \sum_{n=-N}^{N} b_i(n\omega_0)\exp(jn\omega_0 t).$$

We limit ourselves for the time being to unmodulated RF signals:

In a conventional vector network analyzer (NA) only a relative calibration is required, since only the ratio of reflected to incident waves is sought. With a nonlinear vector network analyzer (NVNA) absolute measurements of the power and phase of multi-harmonic signals are performed and additional absolute calibration steps are required for the NVNA.

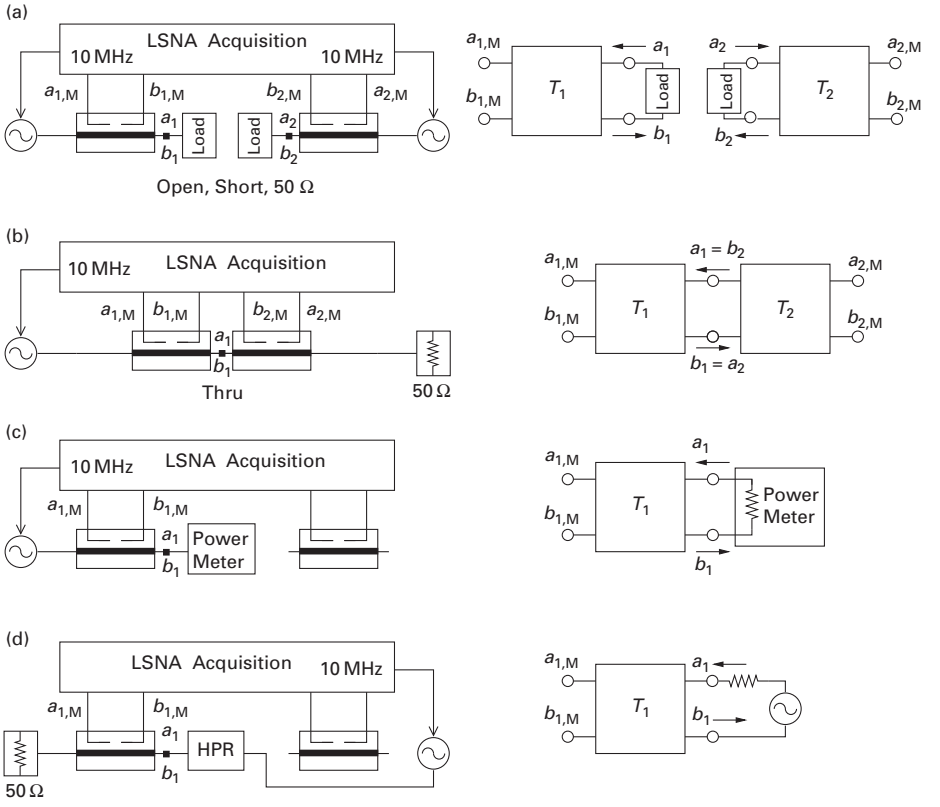


Figure 2.8 Two types of calibration are used: (1) relative calibration: (a) open, short, and load, and (b) thru calibrations; and (2) absolute calibration: (c) power and (d) harmonic phase calibrations.

The major assumption made for calibrating the NVNA is that it is operating itself in a linear regime of operation. This assumption certainly holds true for the couplers and bias tees, assuming that negligible self-heating is taking place. As far as the samplers are concerned, attenuators need to be engaged as needed to make sure that the signal remains in the required dynamic range of the NVNA. Under this linear approximation the waves $a_{i,\mathrm{M}}(n\omega_0)$ and $b_{i,\mathrm{M}}(n\omega_0)$ measured by the NVNA are related by a linear relationship to the waves $a_i(n\omega_0)$ and $b_i(n\omega_0)$ at the reference plane of port i of the DUT. Assuming that channels 1 and 2 for port 1 are fully isolated from channels 3 and 4 of port 2, we obtain the relation

$$\begin{bmatrix} a_{1,\mathrm{M}}(\omega_n) \\ b_{1,\mathrm{M}}(\omega_n) \\ a_{2,\mathrm{M}}(\omega_n) \\ b_{2,\mathrm{M}}(\omega_n) \end{bmatrix} = \begin{bmatrix} A_1(\omega_n) & B_1(\omega_n) & 0 & 0 \\ C_1(\omega_n) & D_1(\omega_n) & 0 & 0 \\ 0 & 0 & A_2(\omega_n) & B_2(\omega_n) \\ 0 & 0 & C_2(\omega_n) & D_2(\omega_n) \end{bmatrix} \begin{bmatrix} a_1(\omega_n) \\ b_1(\omega_n) \\ a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix}, \quad (2.1)$$

where we used the notation $\omega_n = n\omega_0$ for $n \geq 1$.

Since ports 1 and 2 are assumed isolated, let us simply use 2×2 matrices:

$$\begin{bmatrix} a_{1,\mathrm{M}}(\omega_n) \\ b_{1,\mathrm{M}}(\omega_n) \end{bmatrix} = \begin{bmatrix} A_1(\omega_n) & B_1(\omega_n) \\ C_1(\omega_n) & D_1(\omega_n) \end{bmatrix} \begin{bmatrix} a_1(\omega_n) \\ b_1(\omega_n) \end{bmatrix} = \mathbf{T}_1 \begin{bmatrix} a_1(\omega_n) \\ b_1(\omega_n) \end{bmatrix} = A_1(\omega_n)\overline{\mathbf{T}}_1 \begin{bmatrix} a_1(\omega_n) \\ b_1(\omega_n) \end{bmatrix},$$

$$\begin{bmatrix} a_{2,\mathrm{M}}(\omega_n) \\ b_{2,\mathrm{M}}(\omega_n) \end{bmatrix} = \begin{bmatrix} A_2(\omega_n) & B_2(\omega_n) \\ C_2(\omega_n) & D_2(\omega_n) \end{bmatrix} \begin{bmatrix} a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix} = \mathbf{T}_2 \begin{bmatrix} a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix} = A_2\overline{\mathbf{T}}_2 \begin{bmatrix} a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix}.$$

The normalized matrices $\overline{\mathbf{T}}_1(\omega_n)$ and $\overline{\mathbf{T}}_2(\omega_n)$, given by

$$\overline{\mathbf{T}}_i(\omega_n) = \begin{bmatrix} 1 & B_i'(\omega_n) \\ C_i'(\omega_n) & D_i'(\omega_n) \end{bmatrix} = \begin{bmatrix} 1 & B_i(\omega_n)/A_i(\omega_n) \\ C_i(\omega_n)/A_i(\omega_n) & D_i(\omega_n)/A_i(n\omega_n) \end{bmatrix},$$

can be extracted by a relative calibration. This relative calibration can be performed, for example, by using short, open, and load (SOL) standards at ports 1 and 2. Generally speaking, using three such different loads we obtain the system:

$$\begin{bmatrix} -\Gamma_{\mathrm{L}i}(\mathrm{Short}) & \left.\frac{a_{\mathrm{M},i}(\omega_n)}{b_{\mathrm{M},i}(\omega_n)}\right|_{\mathrm{Short}} & \Gamma_{\mathrm{L}i}(\mathrm{Short}) \times \left.\frac{a_{\mathrm{M},i}(\omega_n)}{b_{\mathrm{M},i}(\omega_n)}\right|_{\mathrm{Short}} \\ -\Gamma_{\mathrm{L}i}(\mathrm{Open}) & \left.\frac{a_{\mathrm{M},i}(\omega_n)}{b_{\mathrm{M},i}(\omega_n)}\right|_{\mathrm{Open}} & \Gamma_{\mathrm{L}i}(\mathrm{Open}) \times \left.\frac{a_{\mathrm{M},i}(\omega_n)}{b_{\mathrm{M},i}(\omega_n)}\right|_{\mathrm{Open}} \\ -\Gamma_{\mathrm{L}i}(\mathrm{Load}) & \left.\frac{a_{\mathrm{M},i}(\omega_n)}{b_{\mathrm{M},i}(\omega_n)}\right|_{\mathrm{Load}} & \Gamma_{\mathrm{L}i}(\mathrm{Load}) \times \left.\frac{a_{\mathrm{M},i}(\omega_n)}{b_{\mathrm{M},i}(\omega_n)}\right|_{\mathrm{Load}} \end{bmatrix} \begin{bmatrix} B_i'(\omega_n) \\ C_i'(\omega_n) \\ D_i'(\omega_n) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

which is readily solved for the harmonics $\omega_n = n\omega_0$. For ideal loads this system of equations simplifies as we have $\Gamma_{\mathrm{L}}(\mathrm{Short}) = -1$, $\Gamma_{\mathrm{L}}(\mathrm{Open}) = 1$, and $\Gamma_{\mathrm{L}}(\mathrm{Load}) = 0$.

The ratio between $A_1(\omega_n)$ and $A_2(\omega_n)$ can next be extracted using a thru measurement. A thru admits the following transmission matrix:

$$\begin{bmatrix} a_1(\omega_n) \\ b_1(\omega_n) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix} = \overline{\mathbf{T}}_{\mathrm{thru}} \begin{bmatrix} a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix}.$$

We can then extract the ratio of $A_1(\omega_n)$ and $A_2(\omega_n)$ using the following identity:

$$\begin{bmatrix} a_{1,\mathrm{M}}(\omega_n) \\ b_{1,\mathrm{M}}(\omega_n) \end{bmatrix} = \frac{A_1(\omega_n)}{A_2(\omega_n)} \overline{\mathbf{T}}_1(\omega_n)\overline{\mathbf{T}}_{\mathrm{thru}}\overline{\mathbf{T}}_2^{-1}(\omega_n) \begin{bmatrix} a_{2,\mathrm{M}}(\omega_n) \\ b_{2,\mathrm{M}}(\omega_n) \end{bmatrix}.$$

At this point we have one remaining unknown, say $A_1(\omega_n)$, since we can now express $A_2(\omega_n)$ in terms of $A_1(\omega_n)$. Additional absolute calibrations now need to be performed at port 1 to acquire the amplitude and phase of $A_1(\omega_n)$ for all harmonics ω_n. A calibrated harmonic generator can be used for this purpose. Because the measurement of power is a well-established science, and precise and accurate power meters traceable to world bureaus of standards are readily available, it is common in practice to acquire the amplitude and phase of A_1 separately. A power meter is used at port 1 to perform a power calibration for acquiring the amplitude of $A_1(\omega_n)$.

The system of equations describing this measurement is

$$\begin{aligned} a_{1,\mathrm{M}}(\omega_n) &= A_1(\omega_n)a_1(\omega_n) + B_1(\omega_n)b_1(\omega_n) \\ &= A_1(\omega_n)a_1(\omega_n)\left[1 + B_1'(\omega_n)\Gamma_{\mathrm{P}}(\omega_n)\right], \\ b_{1,\mathrm{M}}(\omega_n) &= C_1(\omega_n)a_1(\omega_n) + D_1(\omega_n)b_1(\omega_n) \\ &= A_1(\omega_n)a_1(\omega_n)\left[C_1'(\omega_n) + D_1'(\omega)\Gamma_{\mathrm{P}}(\omega_n)\right], \\ P_1(\omega_n) &= \frac{1}{2}|a_1(\omega_n)|^2 - \frac{1}{2}|b_1(\omega_n)|^2 \\ &= \frac{1}{2}|a_1(\omega_n)|^2\left(1 - |\Gamma_{\mathrm{L}}(\omega_n)|^2\right), \end{aligned}$$

with $P_1(n\omega_0)$ the power measured at port 1. From the ratio

$$\frac{a_{1,\mathrm{M}}(\omega_n)}{b_{1,\mathrm{M}}(\omega_n)} = \frac{1 + B_1'(\omega_n)\Gamma_{\mathrm{P}}(\omega_n)}{C_1'(\omega_n) + D_1'(\omega_n)\Gamma_{\mathrm{P}}(\omega_n)},$$

we easily extract the reflection coefficient of the power meter for each harmonic:

$$\Gamma_{\mathrm{P}}(n\omega_0) = \frac{b_{1,\mathrm{M}}(\omega_n) - C_1'(\omega_n)a_{1,\mathrm{M}}(\omega_n)}{D_1'(\omega_n)a_{1,\mathrm{M}}(\omega_n) - B_1'(\omega_n)b_{1,\mathrm{M}}(\omega_n)}.$$

The amplitude of $A_1(\omega_n)$ is then obtained from

$$\begin{aligned} |A_1(n\omega_0)| &= \frac{|a_{1,\mathrm{M}}|}{a_1(\omega_n) + B_1'(\omega_n)b_1(\omega_n)} = \frac{|a_{1,\mathrm{M}}(\omega_n)|}{|a_1(\omega_n)|\left[1 + B_1'(\omega_n)\Gamma_{\mathrm{L}}(\omega_n)\right]} \\ &= \frac{|a_{1,\mathrm{M}}(\omega_n)|}{1 + B_1'(\omega_n)\Gamma_{\mathrm{P}}(\omega_n)}\sqrt{\frac{1 - |\Gamma_{\mathrm{P}}(\omega_n)|^2}{2P_1(\omega_n)}}. \end{aligned}$$

The next calibration step is then to acquire the phase of each harmonic. Note that the phase of the fundamental frequency is arbitrary, as the NVNA measurements are typically started at an arbitrary time. We are therefore interested in the relative phase of the harmonic tones compared with the fundamental tone.

Various types of calibrated generators are available for the phase calibration. They are usually referred to as harmonic phase references (HPRs) or harmonic phase standards (HPSs). They usually consist of a comb generator with wide bandwidth (20 to 50 GHz). They can be implemented using an impulse generator to generate a large number of harmonic frequencies (the comb) covering the bandwidth of the NVNA.

A step recovery diode is typically used for generating the impulse (see Figure 2.9). A nonlinear transmission line can also be used for decreasing the impulse rising time. Alternatively, a monolithic microwave integrated circuit (MMIC) can be used to generate the required pulse, such as in the U9391C Agilent module. The key feature of the HPR is that the phase relationships between all the tones be stable.

Let us now discuss how one can define a phase relationship between the harmonic tone $n\omega_0$ and the fundamental ω_0. Given that enough padding (attenuation) is added inside the HPR, a linear model can be used for the signal generated by the HPR at port 1:

$$a_1(n\omega_0) = b_{\mathrm{R}}(n\omega_0) + \Gamma_{\mathrm{R}}(n\omega_0)b_1(n\omega_0). \tag{2.2}$$

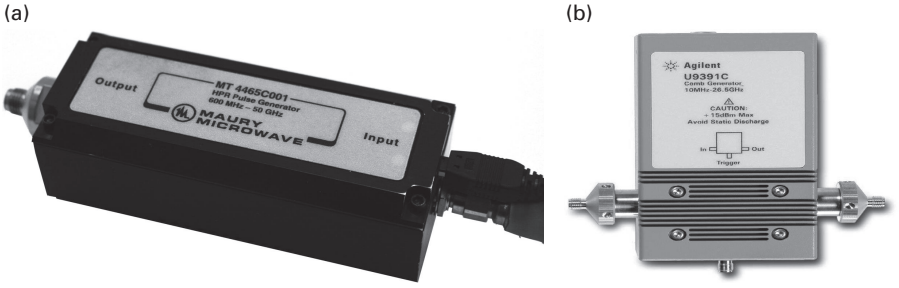


Figure 2.9 (a) Harmonic phase reference/standard from NMDG/Maury (MT4465, 600 MHz to 50 GHz) and (b) comb generator from Agilent (U9391C, 10 MHz to 26.5 GHz) (reproduced with permission from Agilent Technologies, Inc.).

The terms $b_{\mathrm{R}}(n\omega_0) = |b_{\mathrm{R},n}|\exp[j\phi(n,1)]$ are the harmonic components of the periodic signal $b_{\mathrm{R}}(t)$ generated by the HPR,

$$\begin{aligned} b_{\mathrm{R}}(t) = \mathrm{Re}\big\{ & |b_{\mathrm{R},1}|\exp(j\omega_0 t) \\ & + |b_{\mathrm{R},2}|\exp\big(j2\omega_0 t + j\phi_{2,1}\big) \\ & + |b_{\mathrm{R},3}|\exp\big(j3\omega_0 t + j\phi_{3,1}\big) \\ & + \cdots \\ & + |b_{\mathrm{R},N}|\exp\big(jN\omega_0 t + j\phi_{N,1}\big)\big\}, \end{aligned}$$

in the *natural* time reference ($t = 0$) for which the phase of the fundamental is 0 (e.g. $\phi_{1,1} = 0$). Assume now that we measure the HPR signal $b_{\mathrm{R}}(t)$ at a time t_{M} (unknown to us) relative to the *natural time reference of the HPR*, defining thus a new time reference $t' = t - t_{\mathrm{M}}$. The signal measured is then

$$\begin{aligned} b_{\mathrm{R}}(t') = \mathrm{Re}\big\{ & |b_{\mathrm{R},1}|\exp\big[j\omega_0(t' + t_{\mathrm{M}})\big] \\ & + |b_{\mathrm{R},2}|\exp\big[j2\omega_0(t' + t_{\mathrm{M}}) + j\phi_{2,1}\big] \\ & + |b_{\mathrm{R},3}|\exp\big[j3\omega_0(t' + t_{\mathrm{M}}) + j\phi_{3,1}\big] \\ & + \cdots \\ & + |b_{\mathrm{R},N}|\exp\big[jN\omega_0(t' + t_{\mathrm{M}}) + j\phi_{N,1}\big]\big\}. \end{aligned}$$

So in the new time reference t' the HPR signal measured is

$$\begin{aligned} b_{\mathrm{R}}(t') = \mathrm{Re}\big\{ & |b_{\mathrm{R},1}|\exp\big(j\omega_0 t' + j\phi_1'\big) \\ & + |b_{\mathrm{R},2}|\exp\big(j2\omega_0 t' + j\phi_2'\big) \\ & + |b_{\mathrm{R},3}|\exp\big(j3\omega_0 t' + j\phi_3'\big) \\ & + \cdots \\ & + |b_{\mathrm{R},N}|\exp\big(jN\omega_0 t' + j\phi_N'\big)\big\}, \end{aligned}$$

with

$$\begin{aligned}\phi'_1 &= \omega_0 t_{\mathrm{M}},\\ \phi'_2 &= \phi_{2,1} + 2\omega_0 t_{\mathrm{M}},\\ \phi'_3 &= \phi_{3,1} + 3\omega_0 t_{\mathrm{M}},\\ &\dots\\ \phi'_N &= \phi_{N,1} + N\omega_0 t_{\mathrm{M}}.\end{aligned}$$

Clearly, from the measurement with a calibrated scope [12] of the HPR output at an arbitrary and unknown time t_{M} we can recover the relative phase:

$$\phi_{n,1} = \phi'_n - n\omega_0 t_{\mathrm{M}} = \phi'_n - n \times \phi'_1.$$

Let us now use this reference HPR signal for extracting the phase $\phi_{A_1}(n\omega_0)$ of $A_1(n\omega_0)$ of the transmission matrix T_1. $A_1(n\omega_0)$ is a complex number that can be expanded in amplitude and phase:

$$A_1(n\omega_0) = |A_1(n\omega_0)| \times \exp\left[j\phi_{A_1}(n\omega_0)\right].$$

We connect the HPR as indicated in Figure 2.8(d). Given the measured $a_{1,\mathrm{M}}$ and $b_{1,\mathrm{M}}$ waves, we can recover the waves $a_1(\omega_n)$ and $b_1(\omega_n)$ at port 1 using

$$\begin{aligned}\begin{bmatrix} a_1(\omega_n)\\ b_1(\omega_n)\end{bmatrix} &= \mathbf{T}_1^{-1}(\omega_n)\begin{bmatrix} a_{1,\mathrm{M}}(\omega_n)\\ b_{1,\mathrm{M}}(\omega_n)\end{bmatrix}\\ &= \frac{1}{|A_1(\omega_n)|\exp\left[j\phi_{A_1}(\omega_n)\right]}\overline{\mathbf{T}}_1^{-1}(\omega_n)\begin{bmatrix} a_{1,\mathrm{M}}(\omega_n)\\ b_{1,\mathrm{M}}(\omega_n)\end{bmatrix}\\ &= \frac{1}{\exp\left[j\phi_{A_1}(\omega_n)\right]}\begin{bmatrix} a'_{1,\mathrm{M}}(\omega_n)\\ b'_{1,\mathrm{M}}(\omega_n)\end{bmatrix},\end{aligned}$$

with

$$\begin{bmatrix} a'_{1,\mathrm{M}}(\omega_n)\\ b'_{1,\mathrm{M}}(\omega_n)\end{bmatrix} = \frac{1}{|A_1(\omega_n)|}\overline{\mathbf{T}}_1^{-1}(\omega_n)\begin{bmatrix} a_{1,\mathrm{M}}(\omega_n)\\ b_{1,\mathrm{M}}(\omega_n)\end{bmatrix}.$$

The HPR harmonic component $b_{\mathrm{R},n}$ defined in the natural time reference t can then be expressed using Equation (2.2), in terms of the $a_1(n\omega_0)$ and $b_1(n\omega_0)$ waves measured at time t_{M} in the time reference t', using

$$\begin{aligned}b_{\mathrm{R},n}\exp(jn\omega_0 t) &= \exp[jn\omega_0(t-t_{\mathrm{M}})]\left[a_1(n\omega_0) - \Gamma_{\mathrm{R}}(n\omega_0)b_1(n\omega_0)\right],\\ |b_{\mathrm{R},n}|\exp(jn\omega_0 t + j\phi_{n,1}) &= \frac{\exp\left[jn\omega_0(t-t_{\mathrm{M}})\right]}{\exp\left[j\phi_{A_1}(n\omega_0)\right]}\left[a'_{1,\mathrm{M}}(n\omega_0) - \Gamma_{\mathrm{R}}b'_{1,\mathrm{M}}(n\omega_0)\right]\\ &= \frac{\exp\left[jn\omega_0(t-t_{\mathrm{M}})\right]}{\exp\left[j\phi_{A_1}(n\omega_0)\right]}b'_{\mathrm{R,M}}(n\omega_0),\end{aligned}$$

with

$$b'_{\mathrm{R,M}}(n\omega_0) = a'_{1,\mathrm{M}}(n\omega_0) - \Gamma_{\mathrm{R}}(n\omega_0)b'_{1,\mathrm{M}}(n\omega_0).$$

Note that the HPR harmonic coefficients $b_{\mathrm{R}}(n\omega_0) = |b_{\mathrm{R},n}|\exp[j\phi(n,1)]$ are time-invariant because they are defined in the natural time reference t, whereas the complex coefficients $a'_{1,\mathrm{M}}(n\omega_0)$ and $b'_{1,\mathrm{M}}(n\omega_0)$ of the measured HPR signal which are defined in the measurement reference time $t' = t - t_{\mathrm{M}}$ will vary from measurement to measurement as the measurement starting time t_{M} varies arbitrarily.

According to the above expression, the following phase relation is obtained:

$$\phi_{A_1}(n\omega_0) = -\phi_{n,1} - n \times \omega_0 t_{\mathrm{M}} + \text{phase}\left\{b'_{\mathrm{R,M}}(n\omega_0)\right\}.$$

For the fundamental frequencies this reduces to the following relation:

$$\phi_{A_1}(\omega_0) = -\omega_0 t_{\mathrm{M}} + \text{phase}\left\{b'_{\mathrm{R,M}}(\omega_0)\right\}. \tag{2.3}$$

Since again the LSNA measurement of the HPR starts at a random time t_{M} unknown to the user, the phase $\phi_{A_1}(\omega_0)$ cannot be determined (one equation, two unknowns). However, the phase $\phi_{A_1}(\omega_0)$ should remain constant from measurement to measurement since the phase shift from port 1 to the LSNA samplers via the coupler is not time-varying. It is now convenient to select $\phi_{A_1}(\omega_0)$ to be zero since we know it is constant and yet we cannot measure it. This assumption is equivalent to defining a user-defined calibration time reference $t_{\mathrm{M}}(\mathrm{cal})$ obtained by setting $\phi_{A_1}(\omega_0) = 0$ in Equation (2.3):

$$\omega_0 t_{\mathrm{M}}(\mathrm{cal}) = \text{phase}\left\{b'_{\mathrm{R,M}}(\omega_0)\right\}.$$

This time reference $t_{\mathrm{M}}(\mathrm{cal})$ is the correct value t_{M} (modulo $2\pi/\omega_0$) up to an unknown but constant delay:

$$\Delta t_{\mathrm{M}} = t_{\mathrm{M}} - t_{\mathrm{M}}(\mathrm{cal}) = -\phi_{A,1}/\omega_0,$$

within an RF period.

Using this user-defined calibration time reference $t_{\mathrm{M}}(\mathrm{cal})$ the relative phase calibration of the remaining harmonics follows:

$$\phi_{A_1}(n\omega_0) = \text{phase}\left\{b'_{\mathrm{R,M}}(n\omega_0)\right\} - \phi_{n,1} - n \times \text{phase}\left\{b'_{\mathrm{R,M}}(\omega_0)\right\}.$$

The phase of the harmonics will then be correctly set relative to the fundamental frequency, even though the phase of the fundamental frequency will remain arbitrary. It is indeed the very intent of the HPR calibration to establish the correct phase relationship between the harmonics and the fundamental tone.

2.4.2 On-wafer calibration

We have discussed in the previous section the calibration of the LSNA for the measurement of connectorized devices. The absolute calibration steps relied on the use of a power meter and a harmonic phase reference (HPR) at port 1. However, for on-wafer measurements, power meters or HPR generators calibrated at the probe contacts are not available and an alternative approach must be devised. An auxiliary port (port 3) is then used to access the port-1 reference plane on the wafer as shown on Figure 2.10. This

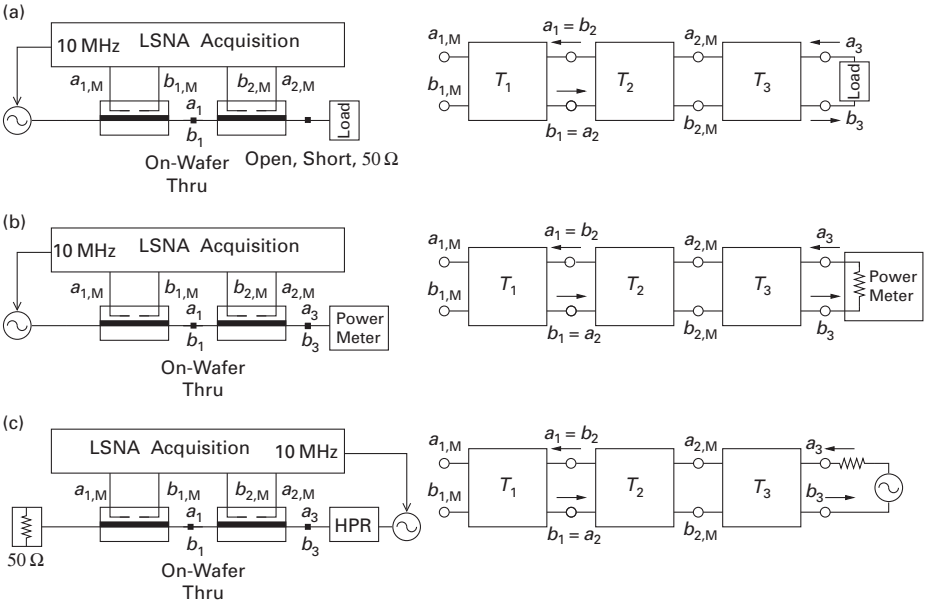


Figure 2.10 Absolute calibration for on-wafer measurements: (a) auxiliary port, (b) power, and (c) HRP calibrations.

can be done easily via an on-wafer thru between port 1 and port 2, and the coupled port of the load coupler (port 2).

One can then recover the signal at port 1 if the transmission matrix T_{23} from port 3 to port 2 is fully characterized. To extract the transmission matrix $T_3(\omega_n)$ between port 3 and the NVNA measurement port 2M, a SOL (short, open, and load) calibration is first performed. A SOL calibration will characterize only the normalized matrix $\overline{T_3}(\omega_n)$, leaving $A_3(\omega_n)$ to be characterized separately:

$$\begin{bmatrix} a_3^{\mathrm{M}}(\omega_n) \\ b_3^{\mathrm{M}}(\omega_n) \end{bmatrix} = \begin{bmatrix} A_3(\omega_n) & B_3(\omega_n) \\ C_3(\omega_n) & D_3(\omega_n) \end{bmatrix} \begin{bmatrix} a_3(\omega_n) \\ b_3(\omega_n) \end{bmatrix}$$
$$= \mathbf{T}_3 \begin{bmatrix} a_3(\omega_n) \\ b_3(\omega_n) \end{bmatrix} = A_3(\omega_n)\overline{\mathbf{T}}_3(\omega_n) \begin{bmatrix} a_3(\omega_n) \\ b_3(\omega_n) \end{bmatrix}.$$

It results that the transmission matrix $\overline{\mathbf{T}}_{23}(\omega_n)$ is given by

$$\begin{bmatrix} A_{23}(\omega_n) & B_{23}(\omega_n) \\ C_{23}(\omega_n) & D_{23}(\omega_n) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{A_3(\omega_n)}{A_2(\omega_n)} \overline{\mathbf{T}}_2^{-1}(\omega_n)\overline{\mathbf{T}}_3(\omega_n).$$

The ratio $A_3(\omega_n)/A_2(\omega_n)$, which is unknown, can be obtained using the reciprocity property

$$A_{23}(\omega_n)D_{23}(\omega_n) - C_{23}(\omega_n)B_{23}(\omega_n) = 1.$$

The ratio $R_n^2 = (A_3(\omega_n)/A_2(\omega_n))^2$ is then determined, and the ratio $R_n = A_3(\omega_n)/A_2(\omega_n)$ is then known up to an undetermined sign for each frequency ω_n.

The sign for the harmonic tones ($n > 1$) can be obtained relative to the fundamental tone by unwrapping the phase (removing 2π phase jump) and selecting the sign such that there is no π phase jump:

$$R_n = \sqrt{|R_n|^2}\exp\left[\text{unwrapped\_phase}\left\{R_n^2\right\}/2\right].$$

The HPR measurement can be used for this phase unwrapping. Indeed, the HPR which is usually set to a subharmonic of the fundamental provides closely spaced frequencies for monitoring the continuity of the unwrapped phase (actually minimizing phase jumps).

In this procedure the sign used for recovering R_1 from R_1^2 at the fundamental frequency is unknown. However, as was indicated in the previous section, the only information required from the HPR is that the phase of the harmonics relative to the fundamental be well defined.

We have extracted all of the eight unknown coefficients of matrices $\overline{\mathbf{T}}_1(\omega_n)$ and $\overline{\mathbf{T}}_2(\omega_n)$. By taking the inverse of Equation (2.1), we can now calculate the actual waves $a_i(\omega_n)$ and $b_i(\omega_n)$ at the reference plane of port i from the signal $a_{i,\mathrm{M}}(\omega_n)$ and $b_{i,\mathrm{M}}(\omega_n)$ measured at the sampler using

$$\begin{bmatrix} a_1(\omega_n) \\ b_1(\omega_n) \\ a_2(\omega_n) \\ b_2(\omega_n) \end{bmatrix} = K_{\mathrm{C}}(\omega_n) \begin{bmatrix} 1 & \beta_1(\omega_n) & 0 & 0 \\ \gamma_1(\omega_n) & \delta_1(\omega_n) & 0 & 0 \\ 0 & 0 & \alpha_2(\omega_n) & \beta_2(\omega_n) \\ 0 & 0 & \gamma_2(\omega_n) & \delta_2(\omega_n) \end{bmatrix} \begin{bmatrix} a_{1,\mathrm{M}}(\omega_n) \\ b_{1,\mathrm{M}}(\omega_n) \\ a_{2,\mathrm{M}}(\omega_n) \\ b_{2,\mathrm{M}}(\omega_n) \end{bmatrix}.$$

2.5 Tuner deembedding with the LSNA

NVNAs are beneficially used in conjunction with tuners to measure the response of a device under realistic load and source impedance terminations. The ideal configuration for this purpose is to place the NVNA couplers between the DUT and the tuners as is indicated in Figure 2.11. This calls for very-low-loss couplers [13]. An example of such couplers reported by XLIM is the wave probe [14]. The wave probe exhibits a low insertion loss at the price of reduced directivity. Note that the directivity error is corrected in the calibration. The wave probe also features an increased coupling at high frequency, which is highly desirable since the harmonics are usually much smaller in amplitude. This results in an effective increase in dynamic range for the harmonic components. The wave probes and the alternative current/voltage-probes implementation are now available in commercial tuners.

In some circumstances it is, however, necessary to place the tuners between the DUT and the NVNA couplers. This occurs when the couplers exhibit too much loss and/or are already integrated in the NVNA. Since now the tuners are in the path between the NVNA couplers and the reference ports 1 and 2, the calibration will be dependent on the tuners' setting. It results that when the tuners are adjusted to provide new different impedances, the calibration of the NVNA is no longer applicable to the new settings.

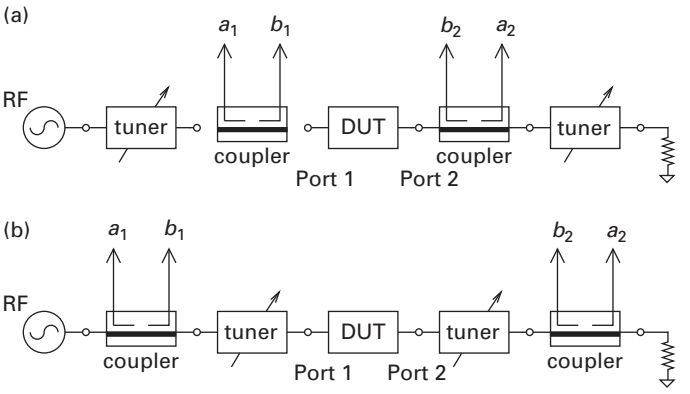


Figure 2.11 Preferred configurations (a) for the integration of tuners with NVNAs and non-optimal integration, and (b) used for lossy couplers.

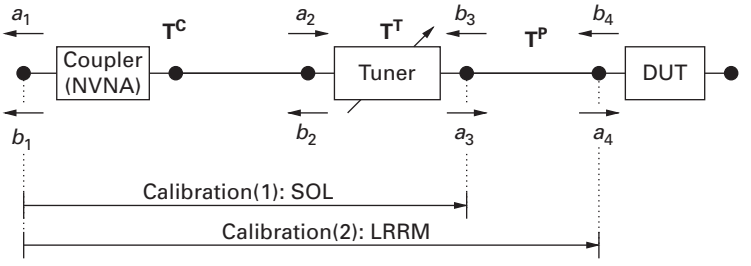


Figure 2.12 A simplified schematic outline of tuner deembedding.

Since the calibration of a network analyzer is a time-consuming process, recalibrating the NVNA each time the tuner is set to a different impedance would be quite impractical. To address this issue a two-tier calibration can be performed [15]. This procedure, described below, relies on the preliminary characterization of the tuner with a network analyzer for the frequency of operation and its harmonics. The computer-controlled tuners also need to exhibit highly reproducible microwave characteristics for this procedure to work.

The two-tier calibration with an NVNA consists of (1) a SOL calibration and (2) a line–reflect–reflect–match (LRRM) calibration. These two calibrations are done with the tuner impedance set to 50 Ω (Z_0). The goal of the tuner deembedding is to recover the correct calibration setup without any extra calibration steps, even if the tuners are placed on new impedance positions (Z_x).

2.5.1 Definitions

Figure 2.12 shows a simplified block diagram of the NVNA measurement system for tuner deembedding. Only one port is considered here.

The SOL calibration is done at the reference plane 3, while the LRRM calibration is done at the reference plane 4, which is generally at the RF probe tips in on-wafer

measurements. Between the reference planes 1 and 3, we can write an error coefficient (ECF) matrix [16] from the NVNA SOL calibration as

$$\begin{bmatrix} a_3 \\ b_3 \end{bmatrix} = K_{\mathrm{S}} \begin{bmatrix} 1 & \beta_{\mathrm{S}} \\ \gamma_{\mathrm{S}} & \delta_{\mathrm{S}} \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}.$$

For clarity of presentation we have dropped the functional dependence on the radial frequency (ω_n) for all the terms in this section. Similarly, for the reference planes 1 and 4, the ECF matrix from the NVNA LRRM calibration can be written as

$$\begin{bmatrix} a_4 \\ b_4 \end{bmatrix} = K_{\mathrm{L}} \begin{bmatrix} 1 & \beta_{\mathrm{L}} \\ \gamma_{\mathrm{L}} & \delta_{\mathrm{L}} \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}.$$

Note that the ECF matrices are sometime stored as the relationship between v and i and the **SOL** or **LRRM** transfer matrices need to be converted for a and b [17].

In addition, the transfer matrix $\mathbf{T}^{\mathrm{T}}$ of the tuner can be obtained from the S-parameters of the tuner [18]:

$$\mathbf{T}^{\mathrm{T}} = \begin{bmatrix} T_{11}^{\mathrm{T}} & T_{12}^{\mathrm{T}} \\ T_{21}^{\mathrm{T}} & T_{22}^{\mathrm{T}} \end{bmatrix} = \begin{pmatrix} 1/S_{21}^{\mathrm{T}} & -S_{22}^{\mathrm{T}}/S_{21}^{\mathrm{T}} \\ S_{11}^{\mathrm{T}}/S_{21}^{\mathrm{T}} & S_{12}^{\mathrm{T}} - S_{11}^{\mathrm{T}} S_{22}^{\mathrm{T}}/S_{21}^{\mathrm{T}} \end{pmatrix}.$$

This matrix is used to relate the reference waves at plane 2 to the reference waves at plane 3:

$$\begin{bmatrix} a_2 \\ b_2 \end{bmatrix} = \mathbf{T}^{\mathrm{T}} \begin{bmatrix} b_3 \\ a_3 \end{bmatrix}. \tag{2.4}$$

The T-matrix of the tuner is subsequently assumed to be known for all possible impedance transformation it provides.

In Figure 2.12 there are two unknown transfer matrices, $\mathbf{T}^{\mathrm{C}}$ and $\mathbf{T}^{\mathrm{P}}$. The matrix $\mathbf{T}^{\mathrm{C}}$ represents the combined characteristics of the NVNA coupler and access cable which is used to connect the NVNA to the tuner (reference plane 2). The matrix $\mathbf{T}^{\mathrm{C}}$ can therefore be represented in an ECF matrix form:

$$\mathbf{T}^{\mathrm{C}} = K_{\mathrm{C}} \begin{bmatrix} 1 & \beta_{\mathrm{C}} \\ \gamma_{\mathrm{C}} & \delta_{\mathrm{C}} \end{bmatrix}. \tag{2.5}$$

This matrix is used to relate the waves at reference plane 2 to the waves at reference plane 1:

$$\begin{bmatrix} a_2 \\ b_2 \end{bmatrix} = \mathbf{T}^{\mathrm{C}} \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}. \tag{2.6}$$

It is also necessary to account for the connection between the tuner and the DUT. $\mathbf{T}^{\mathrm{P}}$ is the T-matrix used to relate the waves at reference plane 3 to the waves at reference plane 4:

$$\begin{bmatrix} a_3 \\ b_3 \end{bmatrix} = \mathbf{T}^{\mathrm{P}} \begin{bmatrix} b_4 \\ a_4 \end{bmatrix}. \tag{2.7}$$

Note that $\mathbf{T}^{\mathrm{C}}$ and $\mathbf{T}^{\mathrm{P}}$ could also include the contribution of any external bias tee located in the path.

Once $\mathbf{T}^{\mathrm{C}}$ and $\mathbf{T}^{\mathrm{P}}$ in Figure 2.12 are available, the modified NVNA calibration can be calculated for any new setting of the tuner. The unknown $\mathbf{T}^{\mathrm{C}}$ and $\mathbf{T}^{\mathrm{P}}$ matrices need therefore to be extracted and this is achieved using the two-tier calibration described in the next section.

2.5.2 Extraction of β_{C}, γ_{C}, and δ_{C} in $\mathbf{T}^{\mathrm{C}}$

The SOL calibration provides us with the normalized T-matrix $\overline{\mathbf{SOL}(Z_0)}$ from the reference plane 1 to plane 3. The SOL calibration matrix $\mathbf{SOL}(\mathbf{Z_0})$ can therefore be expressed in terms of $\mathbf{T}^{\mathrm{C}}$ and $\mathbf{T}^{\mathrm{T}}(Z_0)$ using:

$$\overline{\mathbf{SOL}(Z_0)}^{-1} = \overline{\mathbf{T}^{\mathrm{C}}}^{-1} \times \overline{\mathbf{T}^{\mathrm{T}}(Z_0)}. \tag{2.8}$$

The functional dependence on Z_0 is used to signify that the calibration is performed with the tuner in its through position. The overline used for each matrix T indicates that the matrix is normalized by T_{11}. The calibration matrix $\overline{\mathbf{T}^{\mathrm{C}}}$ is then

$$\begin{aligned}
\overline{\mathbf{T}^{\mathrm{C}}} = \begin{bmatrix} 1 & \beta_{\mathrm{C}} \\ \gamma_{\mathrm{C}} & \delta_{\mathrm{C}} \end{bmatrix} &= \overline{\mathbf{T}^{\mathrm{T}}(\mathbf{Z_0})} \times \overline{\mathbf{SOL}(Z_0)} \\
&= \begin{bmatrix} T_{11}^{\mathrm{T}} & T_{12}^{\mathrm{T}} \\ T_{21}^{\mathrm{T}} & T_{22}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} 1 & \beta_{\mathrm{S}} \\ \gamma_{\mathrm{S}} & \delta_{\mathrm{S}} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \dfrac{T_{11}^{\mathrm{T}}\beta_{\mathrm{S}} + T_{12}^{\mathrm{T}}\delta_{\mathrm{S}}}{T_{11}^{\mathrm{T}} + T_{12}^{\mathrm{T}}\gamma_{\mathrm{S}}} \\ \dfrac{T_{21}^{\mathrm{T}} + T_{22}^{\mathrm{T}}\gamma_{\mathrm{S}}}{T_{11}^{\mathrm{T}} + T_{12}^{\mathrm{T}}\gamma_{\mathrm{S}}} & \dfrac{T_{21}^{\mathrm{T}}\beta_{\mathrm{S}} + T_{22}^{\mathrm{T}}\delta_{\mathrm{S}}}{T_{11}^{\mathrm{T}} + T_{12}^{\mathrm{T}}\gamma_{\mathrm{S}}} \end{bmatrix}.
\end{aligned} \tag{2.9}$$

2.5.3 Extraction of $(1/K_{\mathrm{C}})\mathbf{T}^{\mathrm{P}}$

The LRRM calibration establishes the calibration T-matrix from the reference plane 1 to the reference plane 4. Using Equations (2.4), (2.6), and (2.7), its T-matrix $\mathbf{LRRM}(Z_0)$ can therefore be expressed in the form

$$\mathbf{LRRM}(Z_0)^{-1} = \mathbf{T}^{\mathrm{C}^{-1}} \times \mathbf{T}^{\mathrm{T}}(Z_0) \times \mathbf{T}^{\mathrm{P}}. \tag{2.10}$$

On renormalizing $\mathbf{T}^{\mathrm{C}}$ by $T_{11}^{\mathrm{C}} = K_{\mathrm{C}}$, we obtain

$$\mathbf{LRRM}(Z_0)^{-1} = \overline{\mathbf{T}^{\mathrm{C}}}^{-1} \times \mathbf{T}^{\mathrm{T}}(Z_0) \times (1/K_{\mathrm{C}})\mathbf{T}^{\mathrm{P}}. \tag{2.11}$$

It results that the product of the normalization factor $(1/K_{\mathrm{C}})$ and the transfer matrix $\mathbf{T}^{\mathrm{P}}$ is given by

$$(1/K_{\mathrm{C}})\mathbf{T}^{\mathrm{P}} = \mathbf{T}^{\mathrm{T}}(Z_0)^{-1} \times \overline{\mathbf{T}^{\mathrm{C}}} \times \mathbf{LRRM}(Z_0)^{-1}. \tag{2.12}$$

2.5.4 Extraction of $\mathbf{LRRM}(Z_x)$

Let $\mathbf{T}^{\mathrm{T}}(Z_x)$ be the new transfer matrix for a new tuner position providing an impedance Z_x at the reference plane 3. Then the new LRRM calibration matrix $\mathbf{LRRM}(Z_x)$ for the NVNA accounting for the tuner deembedding can be obtained using the following formula:

$$\mathbf{LRRM}(Z_x) = \left[(1/K_{\mathrm{C}})\mathbf{T}^{\mathrm{P}}\right]^{-1} \times \mathbf{T}^{\mathrm{T}}(Z_x)^{-1} \times \overline{\mathbf{T}^{\mathrm{C}}}, \tag{2.13}$$

where $\overline{\mathbf{T}^{\mathrm{C}}}$ and $(1/K_{\mathrm{C}})\mathbf{T}^{\mathrm{P}}$ are given in Equations (2.9) and (2.12), respectively.

2.6 Modulated measurements and IF calibration

As was mentioned in our discussion of sub-sampling, the sampler-based NVNA is able to down convert not only the fundamental and harmonics but also their modulation in the IF bandwidth of the sampler-based NVNA receiver. It is simply necessary to select the sampling frequency such that no overlap in tones takes place.

However, the issue of what calibration scheme to use for these modulated tones arises. A full RF calibration as described in the previous section is an option but requires knowledge of the modulation frequency at calibration. For relatively small bandwidth it is usually acceptable to rely on the RF calibration which has already been performed for the fundamental or harmonic tones. However, such a calibration must then be corrected for the IF response of the bandpass filter placed before the ADCs. Indeed, the modulation tones associated with each harmonic are down converted to different intermediate frequencies than their corresponding harmonics. The frequency dispersion of the IF filter then must be accounted for.

In the LSNA this is achieved by partitioning the calibration matrix into an RF component and an IF component. The RF calibration is assumed to be valid up to the analog samplers and the IF calibration accounts for the path from the analog samplers to the ADCs. Various IF calibration procedures can then be used to extract the IF filter response. A detailed procedure involving the use of a calibrated modulated reference source for the absolute IF phase calibration is provided in Section 2.7. On the user side it is desirable to check the *relative* IF calibration using a modulated RF source with short and through connections by verifying that the following identities hold:

$$\left.\begin{aligned} b_2(n\omega_0 + p\omega_m) &= a_1(n\omega_0 + p\omega_m) \\ b_1(n\omega_0 + p\omega_m) &= a_2(n\omega_0 + p\omega_m) \end{aligned}\right\} \text{ for a through connection,}$$

$$\left.\begin{aligned} b_1(n\omega_0 + p\omega_m) &= -a_1(n\omega_0 + p\omega_m) \\ b_2(n\omega_0 + p\omega_m) &= -a_2(n\omega_0 + p\omega_m) \end{aligned}\right\} \text{ for a short termination,}$$

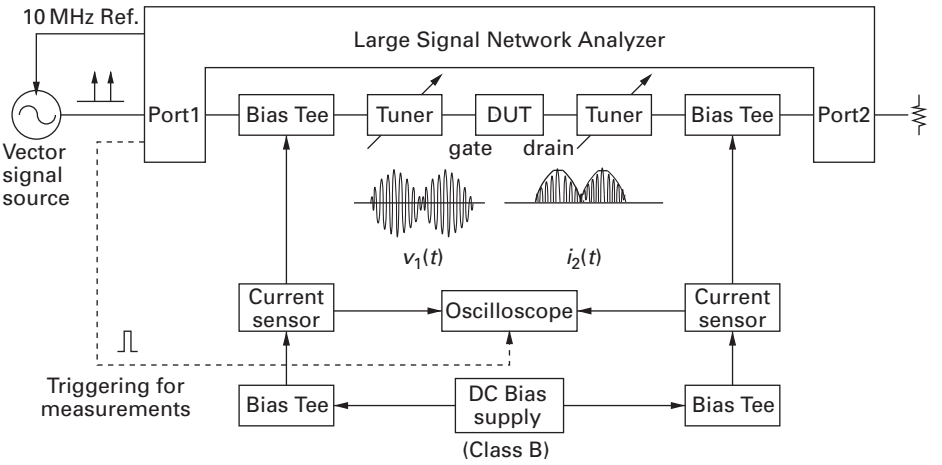


Figure 2.13 Absolute time-reference calibration with a two-tone modulation.

for all tones p and harmonics n. Residual errors observed usually take the form of a group delay, which can easily be removed as long as the errors are linear and deterministic (invariant under successive measurements). Such an example is described in the next section when synchronizing the RF and baseband measurements.

2.6.1 Absolute time reference calibration for RF modulated measurements

The complete characterization of a device excited by modulated RF signal requires the acquisition both of the RF and of the baseband signals. Such a measurement system is shown in Figure 2.13 for the case of an LSNA configured with a trigger and used in conjunction with an oscilloscope.

For CW-RF signals only the DC voltages and currents are to be acquired at baseband and no synchronization issue is involved. But when modulation is involved, it becomes critical to attain a good synchronization of the RF and baseband measurement systems. Indeed, the sampler-based NVNA used for the RF measurements and the oscilloscope used for the baseband signal measurements might start their data acquisition at different times, even though they share the same starting trigger signals. This originates among other things from the delay produced by the different cable lengths used in the RF path and the current sensor dispersion in the baseband path.

A timing calibration can be achieved by using a two-tone measurement for a fast reference transistor operated under class-B biasing conditions. In class-B operation, the DUT acts as an RF signal detector at port 2. With the testbed calibrated, a two-tone signal is applied to the gate input of the DUT:

$$v_1(t) = \cos(\omega_1 t) + \cos(\omega_2 t). \tag{2.14}$$

The transistor output in class B neglecting capacitance and assuming linear amplification is then

$$
\begin{aligned}
i_2(t) &= G_{\mathrm{m}} u \left\{\cos(w_1 t + \phi_1) + \cos(\omega_2 t + \phi_2)\right\} \\
&= G_{\mathrm{m}} u \left\{\cos\left[\frac{1}{2}(\omega_2 - \omega_1)t + \frac{1}{2}(\phi_2 - \phi_1)\right]\right. \\
&\qquad \left.\times \cos\left[\frac{1}{2}(\omega_2 + \omega_1)t + \frac{1}{2}(\phi_2 + \phi_1)\right]\right\}, \qquad (2.15)
\end{aligned}
$$

where $u[x]$ is the step function. Equation (2.15) explicitly shows the amplitude modulation at $(\omega_2 - \omega_1)$ induced with the beating of the two-tone excitations. The rectifying function $u[x]$ will down convert this amplitude modulation to baseband. However, the LSNA used in this measurement ignores the signal below 600 MHz, so that the measured RF signal does not account for the baseband tones. It results that the RF output current acquired does not appear rectified even though the DC bias gate voltage is set to the threshold voltage as needed for proper class-B operation. For this reason, the baseband waveforms must be acquired and added to the RF waveforms. As shown in Figure 2.13, this is done using current sensors to measure the baseband current waveforms using an oscilloscope.

The absolute time reference calibration is done by determining the time shift which is required in order to obtain a rectified drain current output after the baseband intermodulation products $I_{\mathrm{DS}}(t)$ of the oscilloscope are added to the RF drain current waveforms $i_2(t)$ of the LSNA.

Figure 2.14 illustrates these IF calibration steps. It is clearly shown in Figure 2.14(a) that the measured RF drain current $i_2(t)$ and the baseband drain current $I_{\mathrm{DS}}(t)$ are not

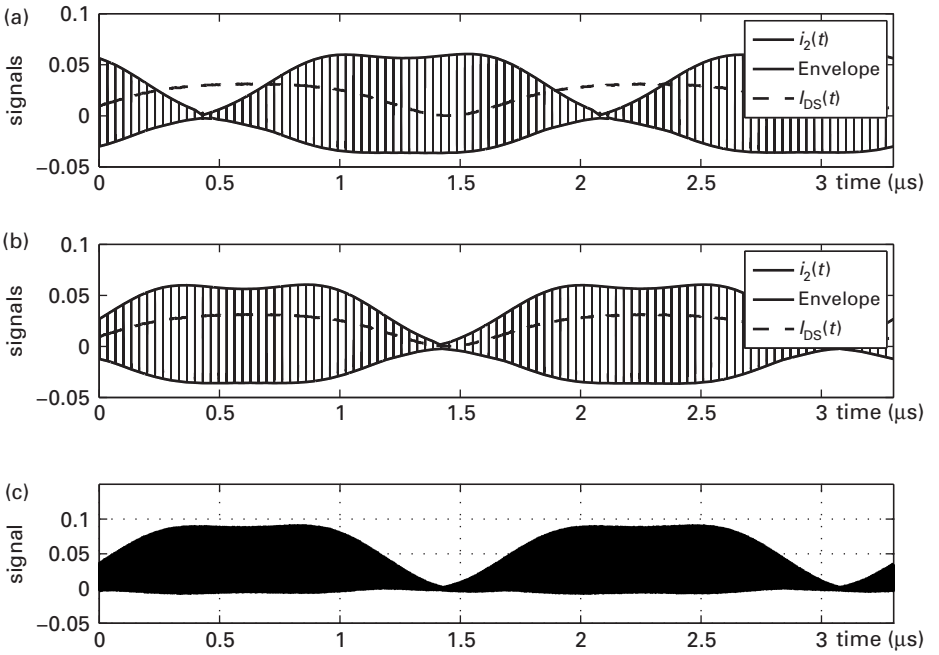


Figure 2.14 Time reference for the IF calibration of the LSNA: (a) measured, (b) synchronized, and (c) added RF and baseband signals.

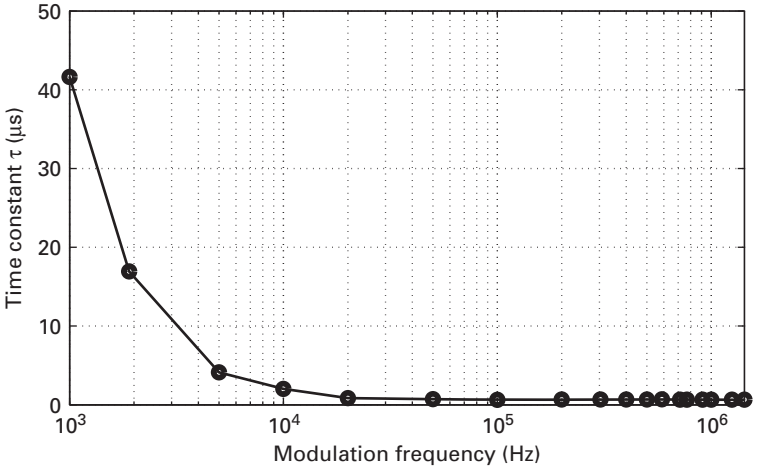


Figure 2.15 Extracted time constants as a function of modulation frequency.

synchronized. In order to synchronize those two signals, it is sufficient to align the time of zero crossing for the baseband $I_{DS}(t)$ and the RF envelope of $i_2(t)$. This is illustrated in Figure 2.14(b). When $i_2(t)$ coincides with $I_{DS}(t)$, the sum of two signals produces the rectified RF drain current shown in Figure 2.14(c), as expected for class-B operation.

In Figure 2.14, the required time constant (τ) is 0.6695 μs for the LSNA modulation frequency (f_m) of 303.07 kHz. Since the time constant is a function of the modulation frequency used, it is necessary to find out the corresponding time constants corresponding to all modulation frequencies interested. The variation of the time constants extracted is illustrated in Figure 2.15. This dispersion below 20 kHz is associated here with the current sensor in the baseband path which relies on a transformer optimally designed for the acquisition of current pulses of microsecond duration.

Another way to achieve accurate time reference calibration could be to use a pulsed-RF signal instead of a two-tone signal. The rectified drain current will detect the presence of the RF signal at the port 2.

2.7 Broadband measurements with the LSNA

The growing demand for broadband wireless communication services is motivating the development of power amplifiers (PAs) operating with increasingly wider bandwidth. Nonlinear vector network analyzers (NVNAs) offer unique possibilities for characterizing the nonlinear behavior of PAs for periodically modulated excitations. As we have seen, the sampler-based NVNA can acquire multitone periodic signals with wide RF bandwidth (up to 50 GHz in the LSNA) and effectively remap them under appropriate conditions within the bandwidth of the IF receiver (10 MHz in Maury's LSNA and 50 MHz in VTD's SWAP). Once down converted the RF tones are filtered with a low-pass IF filter and then digitized using analog-to-digital converters (ADCs). In the

commercialized LSNA, the modulation bandwidth for the RF signals acquired was limited to twice the IF modulation bandwidth, e.g. 20 MHz in the LSNA. It is, however, possible under appropriate conditions to further increase the modulation bandwidth, by choosing the measurements parameters of the LSNA such as the ADC frequency f_{ADC}, the LO frequency f_{LO}, and the FFT size N_{FFT} in such a way that none of the fundamental and harmonic tones folds down to the same IF frequency during down conversion [19]. Wideband multitone signals can be acquired that way after the frequencies measured are descrambled. Note that a technique to avoid the frequency scrambling has been demonstrated for wideband LSNA measurements [20]. However, it requires adding a switch between the FracN and the step recovery diode.

Such a broadband measurement system can be realized with a large-signal network analyzer by adding an additional frequency synthesizer to set the ADC clock and a counter driven by the ADC clock for triggering the measurements (see Figure 2.16). Once calibrated, the modulated RF signals are then acquired in a single LSNA measurement. No stitching of measured data from multiple measurements is therefore required [21] [22] in order to obtain the amplitude and phase of the various tones of the multiharmonic wideband periodically modulated signal acquired. However, as we shall see, the broadband measurement requires a trigger with a frequency set to a sub-harmonic of the greatest common divisor frequency f_{gcd} of all the down-converted IF tones [23]. In the case of IF tones in an arithmetical progression $[f_{IF},\ 2f_{IF},\ 3f_{IF},\ \ldots,\ nf_{IF}]$, the triggering frequency reduces to the fundamental IF frequency f_{IF} used for calibrated broadband measurement [24]. This trigger guarantees that the phase relationship between the tones remains the same in successive measurements.

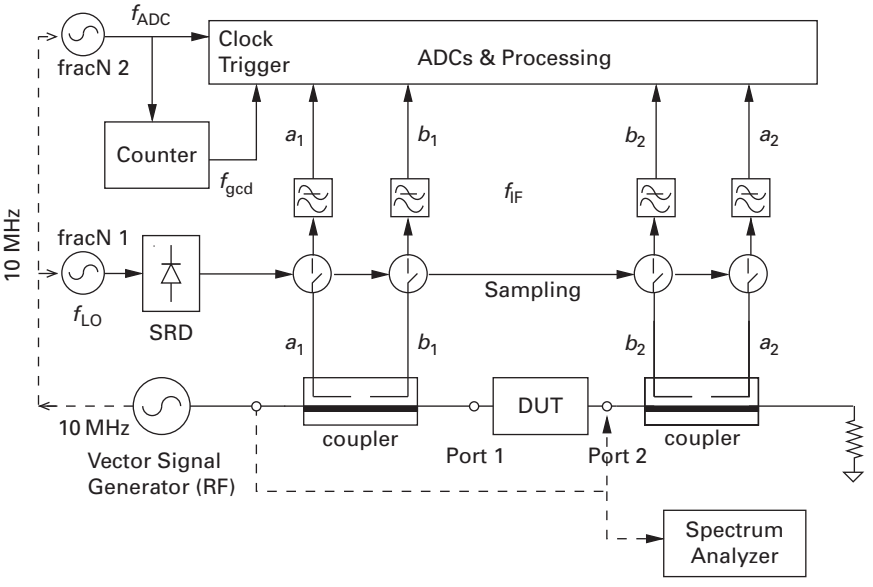


Figure 2.16 Block diagram of the modified LSNA setup and integrated IF calibration setup. (From [23] with permission, ©2010 IEEE.)

2.7.1 Principle of phase calibration

As mentioned above, the bandwidth of the LSNA can be extended beyond the bandwidth of the IF receiver [19] [24] if a broadband calibration scheme is introduced. Consider a periodically modulated signal expressed as the superposition of multiple tones with radial frequency ω_p:

$$y_{\mathrm{RF}} = \sum_{p=1}^{P} A_p \cos(\omega_p t + \phi_p).$$

A_p and ϕ_p are the amplitudes and phases, respectively, for the tones p. The signal y_{RF} being periodically modulated, the radial frequencies ω_p for the tone p can be expressed in terms of the fundamental frequency ω_0 as

$$\omega_p = \omega_0 + p\omega_{\mathrm{m}}, \tag{2.16}$$

where ω_{m} is the modulation frequency corresponding to the tone spacing of the signal y_{RF}.

In the frequency domain, the sub-sampling analog down conversion [19] taking place in the LSNA will down convert each tone of frequency ω_p into the IF bandwidth of the LSNA receiver by mixing it with the harmonics of the comb generator f_{LO}:

$$\begin{aligned} y_{\mathrm{sampler}} &= 2\sum_{p=1}^{P} A_p \cos[\omega_p t + \phi_p] \times \sum_{q=1}^{Q} B_q \cos[q\omega_{\mathrm{LO}} t + \theta_q] \\ &= \sum_{p=1}^{P}\sum_{q=1}^{Q} A_p B_q \big\{\cos[(\omega_p + q\omega_{\mathrm{LO}})t + \phi_p + \theta_q] \\ &\qquad + \cos[(\omega_p - q\omega_{\mathrm{LO}})t + \phi_p - \theta_q]\big\}, \end{aligned} \tag{2.17}$$

where B_q and θ_q are the amplitudes and phases of the harmonics of the comb generator. Note that the signal generator and the local oscillator (comb generator) used are synchronized with respect to a common external reference source (typically 10 or 100 MHz) so that they share the same time base. It results that the phases ϕ_p and θ_q are well defined in the time reference selected once the various generators are powered on. The IF frequency of the tone ω_p after down conversion is

$$\Delta\omega_p = \omega_p - k_p\omega_{\mathrm{LO}} \qquad \text{with} \quad k_p = \mathrm{floor}\left(\frac{\omega_p}{\omega_{\mathrm{LO}}}\right).$$

If we use $q = k_p$ and $q = k_p + 1$ for the second term in (2.17) for each tone p, the multisine signal at the output of the IF filter is then in the frequency range $0 < \omega_p < \omega_{\mathrm{LO}}$:

$$\begin{aligned} y_{\mathrm{IF}}(t) = \sum_{p} & A_p B_{k_p} H_{\mathrm{IF}}(\Delta\omega_p) \cos\left[\Delta\omega_p\, t + \chi_p^{+}\right] \\ & + A_p B_{k_p+1} H_{\mathrm{IF}}(\omega_{\mathrm{LO}} - \Delta\omega_p) \cos\left[\left(\omega_{\mathrm{LO}} - \Delta\omega_p\right) t - \chi_p^{-}\right], \end{aligned} \tag{2.18}$$

with H_{IF} the IF filter amplitude gain,

$$\begin{aligned}\chi_p^+ &= \phi_p - \theta_{k_p} + \varphi_{\mathrm{IF}}[\Delta\omega_p],\\ \chi_p^- &= \phi_p - \theta_{k_p+1} - \varphi_{\mathrm{IF}}[\omega_{\mathrm{LO}} - \Delta\omega_p],\end{aligned} \tag{2.19}$$

and φ_{IF} the IF filter-induced phase shift. Note that it is assumed that the LO frequency ω_{LO} is selected such that we have

$$\frac{\omega_{\mathrm{LO}}}{2} \leq \omega_{\mathrm{IF}},$$

where ω_{IF} is the passband bandwidth of the anti-aliasing IF filter following the sampler. This choice guarantees an LSNA modulation bandwidth of $2\omega_{\mathrm{IF}}$ if we limit the down conversion to a single k_p per harmonic.

Let us assume that the LSNA measurement is now taking place at time $t = \tau_{\mathrm{m}}$. The phases $\psi_p^{\pm(m)}$ of the tones p measured by the LSNA which are defined in the measurement time axis $t' = t - \tau_{\mathrm{m}}$ by

$$\begin{aligned}y_{\mathrm{IF}}(t') = \sum_p & A_p B_{k_p} H_{\mathrm{IF}}(\Delta\omega_p)\cos\left[\Delta\omega_p\, t' + \psi_p^{+(\mathrm{m})}\right]\\ &+ A_p B_{k_p+1} H_{\mathrm{IF}}(\omega_{\mathrm{LO}} - \Delta\omega_p)\cos[(\omega_{\mathrm{LO}} - \Delta\omega_p)t' - \psi^{-(m)}]\end{aligned} \tag{2.20}$$

are then given by

$$\begin{aligned}\psi_p^{+(m)} &= \Delta\omega_p\, \tau_{\mathrm{m}} + \phi_p - \theta_{k_p} + \varphi_{\mathrm{IF}}[\Delta\omega_p],\\ \psi_p^{-(m)} &= -\omega_{\mathrm{LO}}\tau_{\mathrm{m}} + \Delta\omega_p\, \tau_{\mathrm{m}} + \phi_p - \theta_{k_p+1} - \varphi_{\mathrm{IF}}[\omega_{\mathrm{LO}} - \Delta\omega_p],\end{aligned}$$

for positive frequencies $\Delta\omega_p$ and negative (Nyquist image) frequencies $\Delta\omega_p - \omega_{\mathrm{LO}}$, respectively.

It results that the measured phases for both positive and negative frequencies will vary from measurement to measurement since they will take place at different times τ_{m}. Nevertheless to calibrate the LSNA we need the relative phase between the tones to remain constant. Using for the fundamental band the center frequency $\omega_0 = \Delta\omega_0 + k_0\omega_{\mathrm{LO}}$ as the reference, it can readily be verified that the relative phase differences between the tones p and 0 both for positive and for negative (Nyquist image) frequencies are given by

$$\begin{aligned}\psi_p^{+(m)} - \psi_0^{+(m)} &= (\Delta\omega_p - \Delta\omega_0)\tau_{\mathrm{m}} + (\phi_p - \phi_0) - (\theta_{k_p} - \theta_{k_0})\\ &\quad + (\varphi_{\mathrm{IF}}[\Delta\omega_p] - \varphi_{\mathrm{IF}}[\Delta\omega_0]),\\ \psi_p^{-(m)} - \psi_0^{-(m)} &= (\Delta\omega_p - \Delta\omega_0)\tau_{\mathrm{m}} + (\phi_p - \phi_0) - (\theta_{k_p+1} - \theta_{k_0+1})\\ &\quad - (\varphi_{\mathrm{IF}}[\omega_{\mathrm{LO}} - \Delta\omega_p] - \varphi_{\mathrm{IF}}[\omega_{\mathrm{LO}} - \Delta\omega_0]).\end{aligned}$$

The relative phases between the tones are seen to be also measurement-dependent. The measurement-dependent terms are those involving τ_{m}. Using Equation (2.16) we can derive the identity

$$\Delta\omega_p - \Delta\omega_0 = p\omega_{\mathrm{m}} - (k_p - k_0)\omega_{\mathrm{LO}}.$$

The variation of the relative phase differences $\psi_p^{\pm(m)} - \psi_0^{\pm(m)}$ from measurement $m = 2$ to measurement $m = 1$ is then given by

$$\left(\psi_p^{\pm(2)} - \psi_0^{\pm(2)}\right) - \left(\psi_p^{\pm(1)} - \psi_0^{\pm(1)}\right) = p\omega_{\mathrm{m}}\tau_{21} - (k_p - k_0)\omega_{\mathrm{LO}}\tau_{21}, \quad (2.21)$$

with $\tau_{21} = \tau_2 - \tau_1$.

It can be seen from Equation (2.21) that the variation of the relative phase between the p and 0 tones acquired in successive measurements for a given RF harmonic is affected by two group delays: $\omega_{\mathrm{m}}\tau_{21}$ and $\omega_{\mathrm{LO}}\tau_{21}$. The modulation group delay $\omega_{\mathrm{m}}\tau_{21}$ is the usual signal group delay due to the different measurements occurring at different times τ. However, the LO group delay $\omega_{\mathrm{LO}}\tau_{21}$ is a new term that arises in broadband measurement when k_p is different from k_0. In sampler-based NVNAs that are bandwidth-limited at the data-processing level to twice the IF bandwidth ($2\omega_{\mathrm{IF}}$), we have $k_p = k_0$ for each RF harmonic. On the other hand, as we extend the sampler-based NVNA bandwidth to a multiple of twice the IF bandwidth, LO phase shifts $(k_p - k_0)\omega_{\mathrm{LO}}\tau_{21}$ are observed.

If one of the LSNA channels is used to measure a known calibrated signal (e.g. at the PA input) extending over a portion of the spectrum larger than $2\omega_{\mathrm{IF}}$, it is then readily possible to extract both $\omega_{\mathrm{m}}\tau_{21}$ and $\omega_{\mathrm{LO}}\tau_{21}$. This is, however, constraining and requires additional signal processing. The results for the LO delay extraction are shown in Figure 2.17 for three successive calibration measurements. Note that the same LO delay $\omega_{\mathrm{LO}}\tau_{21}$ ($\omega_{\mathrm{LO}}\tau_{31}$) applies to successive bands of $2\omega_{\mathrm{IF}}$ bandwidth.

An alternative approach (see [24]), which is more elegant and convenient, relies on the use of a trigger signal to eliminate altogether the LO group delay $\omega_{\mathrm{LO}}\tau_{21}$ and even the signal group delay $\omega_{\mathrm{m}}\tau_{21}$ (although this is less critical). To synchronize the measurement system, the trigger frequency should be a sub-harmonic of the greatest common divider frequency f_{gcd} for all the IF tones. The following methodology is used for the trigger selection. Let T_{RES} be the LSNA measurement duration during which the ADCs acquire the IF signal. In our measurements (including the calibration) we select the measurement start times to be spaced relatively to each other by a multiple P of T_{RES}:

$$\tau_n = P \times T_{\mathrm{RES}} = P\frac{2\pi}{\omega_{\mathrm{RES}}} \qquad \text{with} \quad \omega_{\mathrm{RES}} = 2\pi f_{\mathrm{RES}} = \frac{2\pi}{T_{\mathrm{RES}}},$$

where $f_{\mathrm{RES}} = 1/T_{\mathrm{RES}}$ is the resolution frequency. Given that the LO frequency ω_{LO}, modulation frequency ω_{m}, and IF frequency $\Delta\omega_0$ of the reference center tone are each selected to be a multiple of the resolution frequency ω_{RES},

$$\omega_{\mathrm{LO}} = N\omega_{\mathrm{RES}}, \qquad \omega_{\mathrm{m}} = Q\omega_{\mathrm{RES}}, \qquad \Delta\omega_0 = K\omega_{\mathrm{RES}},$$

we have for the IF frequencies

$$\begin{aligned} \Delta\omega_p &= \Delta\omega_0 + p\omega_{\mathrm{m}} - (k_p - k_0)\omega_{\mathrm{LO}} \\ &= [K + pQ - (k_p - k_0)N]\omega_{\mathrm{RES}} = M\omega_{\mathrm{RES}}, \\ \omega_{\mathrm{LO}} - \Delta\omega_p &= (N - M)\omega_{\mathrm{RES}}. \end{aligned}$$

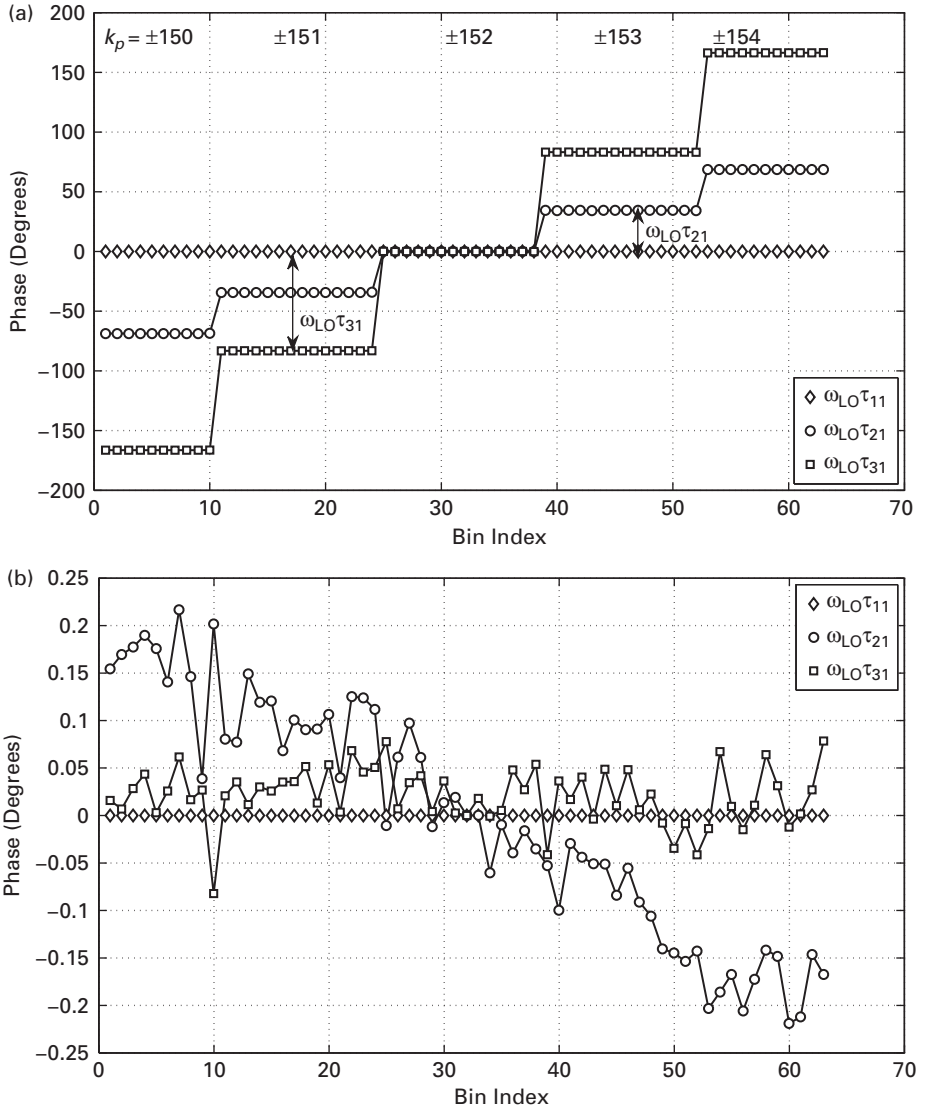


Figure 2.17 Relative LO group delay extraction in both non-trigger (a) and trigger (b) modes. (From [23] with permission, ©2010 IEEE.)

The IF frequencies obtained indicate that f_{RES} is f_{gcd} or a sub-harmonic of it. It further results that

$$\omega_{\mathrm{LO}}\tau_{21} = N\omega_{\mathrm{RES}}P_{21}\frac{2\pi}{\omega_{\mathrm{RES}}} = N \times P_{21} \times 2\pi,$$

$$\omega_{\mathrm{m}}\tau_{21} = Q\omega_{\mathrm{RES}}P_{21}\frac{2\pi}{\omega_{\mathrm{RES}}} = Q \times P_{21} \times 2\pi,$$

$$\Delta\omega_{p}\tau_{21} = M\omega_{\mathrm{RES}}P_{21}\frac{2\pi}{\omega_{\mathrm{RES}}} = M \times P_{21} \times 2\pi,$$

Table 2.1. Signal and LSNA configuration parameters

Parameter	Symbol and value
Center frequency	$f_0 = 2.16\,\text{GHz}$
Number of tones	63
PAPR	6 dB
Modulation frequency	$f_\text{m} = 1.2477\,\text{MHz}$
LO frequency	$f_\text{LO} = 17.361\,975\,\text{MHz}$
ADC frequency	$f_\text{ADC} = 19.6608\,\text{MHz}$
Size of FFT	$N_\text{FFT} = 262\,144$
Resolution frequency	$f_\text{RES} = f_\text{ADC}/N_\text{FFT} = 75\,\text{Hz}$
Trigger frequency	1 Hz
Number of RF harmonics	10
Maximum modulation bandwidth	$f_\text{m} = 1.24\,\text{GHz}$

with $P_{21} = P_2 - P_1$. Clearly, with this choice of measurement times the relative phases between tones extracted by the LSNA are measurement-independent. Indeed, as can be seen in Figure 2.17, the LO group delay between successive measurements exhibits only a small fluctuation in phase thanks to triggering.

The broadband calibration of the sampler-based NVNA proceeds as follows. First a conventional RF calibration including SOLT, power meter, and harmonic phase reference is used for the RF fundamental and harmonics. Then the IF calibration procedure described below is performed at each harmonic band of interest. The "through" IF calibration at each harmonic includes an IF calibration of the incident signal at port 2 and a through calibration between port 1 and port 2. For these IF calibrations a multitone signal with known amplitudes and phases is programmed in a calibrated reference vector signal generator for each harmonic of interest. For the phase calibration, since the relative phases $\psi_p^{+(m)} - \psi_0^{+(m)}$ measured by the sampler-based NVNA are now invariant under the triggered measurements, the phase correction factor ζ_p for the pth tone is then obtained from the reference calibration phase $\psi_p(\text{CAL})$ using

$$\left[\psi_p^{+(m)} - \psi_0^{+(m)}\right] + \zeta_p = \left[\psi_p(\text{CAL}) - \psi_0(\text{CAL})\right].$$

2.7.2 Experimental results and discussions

An example of broadband measurement configuration for the LSNA is given in Table 2.1. It can be verified in this example that no overlap (collision) occurred in the IF bandwidth, for down-converted ω_p tones, ω_m spaced, which are located within a band of bandwidth 1.24 GHz centered at each of the 10 RF harmonics. The experimental verification is, however, done using an 80-MHz vector signal generator.

The phases of a_1 and b_2 at the 63 FFT bins of interest for a 63-tone signal of bandwidth 80 MHz measured with the broadband testbed are shown in Figure 2.18 for a thru after the RF and IF calibration corrections have been applied. Thanks to the properly locked 1-Hz measurement trigger, the phases of a_1 and b_2 at the targeted bins remain

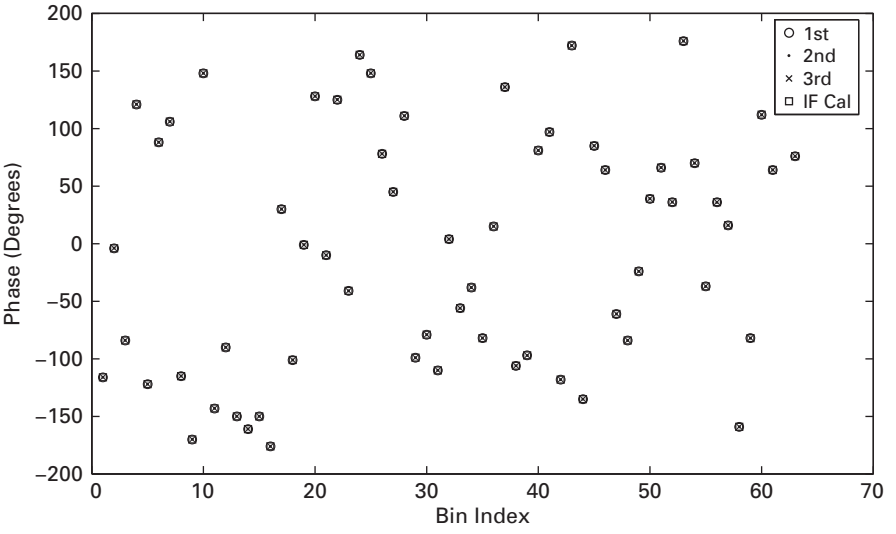


Figure 2.18 The a_1 phase after IF calibration for three successive measurements for a_1. (From [23] with permission, ©2010 IEEE.)

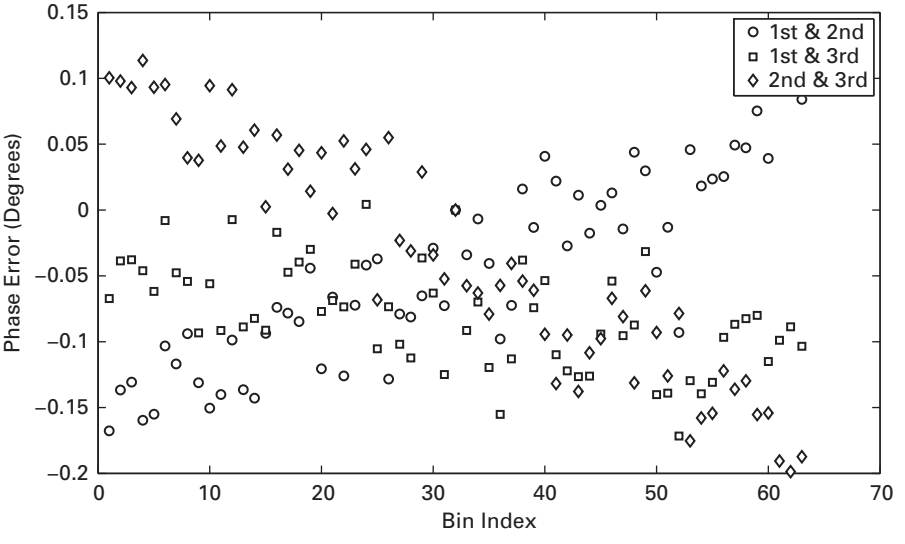


Figure 2.19 The a_1 phase difference among three successive measurements. (From [23] with permission, ©2010 IEEE.)

the same in three successive measurements and correspond to the phases programmed in the vector signal generator. Note that identical phases between a_1 and b_2 result from the fact that a thru is measured. Figure 2.19 illustrates the phase difference observed among three successive measurements. The maximum error range among measurements at each bin indicates (1) that the generated trigger effectively synchronizes the

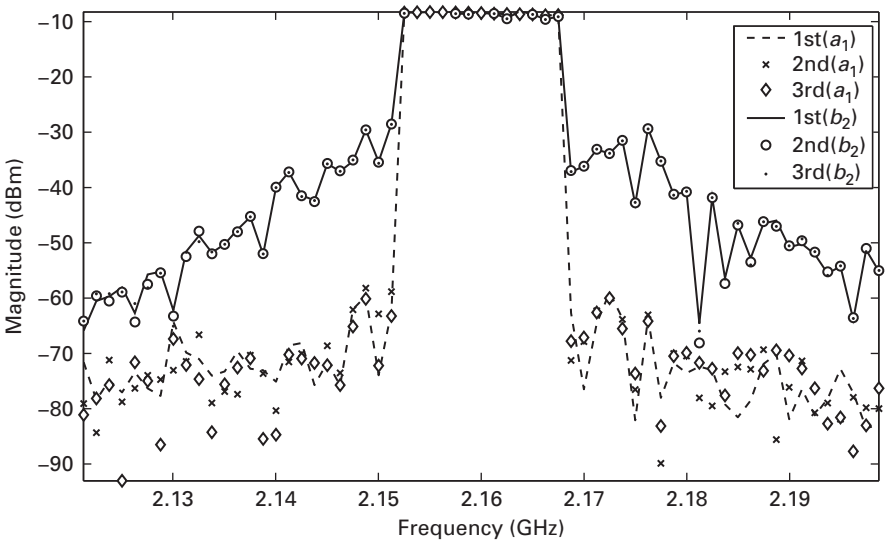


Figure 2.20 Spectral distribution of a_1 and b_2 at the input and output of the PA, respectively, for three successive measurements. (From [23] with permission, ©2010 IEEE.)

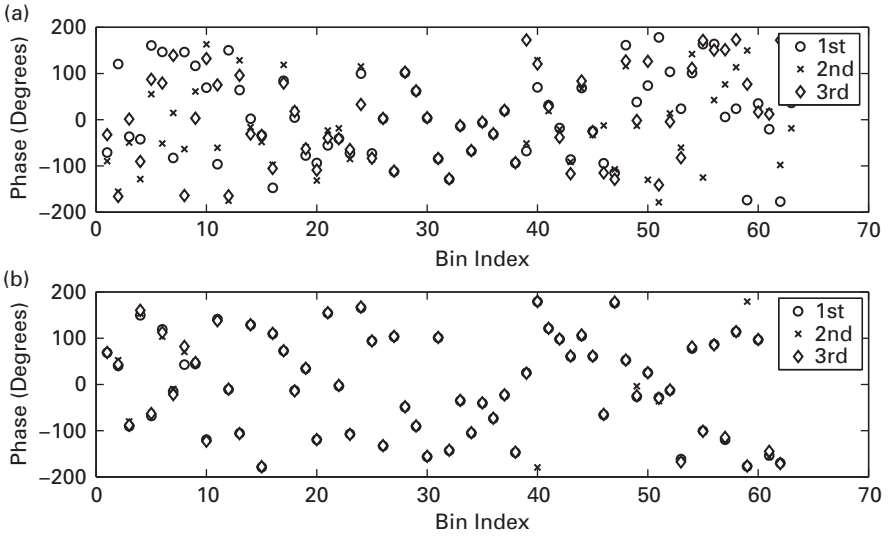


Figure 2.21 Phase distribution of (a) a_1 and (b) b_2 at the input and output of the PA, respectively, for three successive measurements. (From [23] with permission, ©2010 IEEE.)

measurement system (LSNA) and (2) that the vector signal generator used generates a modulated RF signal with stable phases.

Figure 2.20 shows the results obtained for an 80-W GaN HEMT Doherty PA. Displayed in Figures 2.20 and 2.21 are the spectra and phases for the periodic signal at the input (dashed line) and output (plain line) of the PA for three successive measurements. The output spectra are shifted down (by 37 dB) to more clearly reveal the spectral

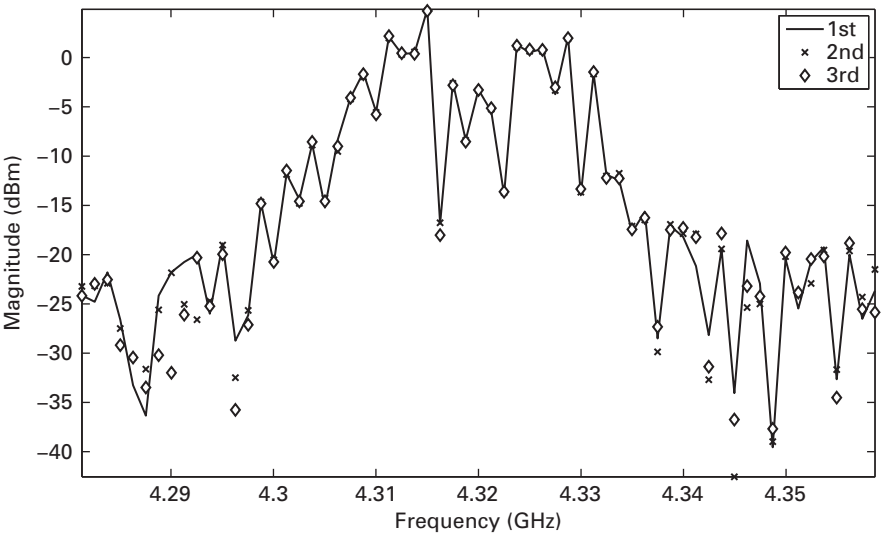


Figure 2.22 Resulting spectral regrowth in the second harmonic at the PA output for three successive measurements. (From [23] with permission, ©2010 IEEE.)

regrowth generated by the PA. The fundamental frequency was set to 2.16 GHz with a multitone bandwidth of around 15 MHz. The number of tones is 13, and the phases are selected to yield a peak-to-average power ratio (PAPR) of 6 dB. The bandwidth of 80 MHz used is wide enough to capture the third- and fifth-order intermodulation bands. As can be seen in Figure 2.20, the spectral regrowth of b_2 is more pronounced than that of a_1. Also, a good phase reproducibility for a_1 for three successive measurements is observed in Figure 2.21 only within the 15-MHz bandwidth of the injected input signal. However, the phase reproducibility for b_2 for three successive measurements is seen in Figure 2.21 to extend over nearly 80 MHz due to the spectral regrowth generated by the PA nonlinearities. The spectral regrowth in the second harmonic of the PA is also presented in Figure 2.22 for three successive measurements. The most reproducible measurements are obtained for tones above −10 dBm.

The error vector magnitude (EVM) is commonly used as a figure of merit for qualifying the in-band linearity performance of a PA for digital communication. For obtaining a highly reliable EVM, the measurement noise error in phase and amplitude contributed by the test system must be smaller than those introduced by the PA. The EVM can also be used to quantify the vector measurement reproducibility error of the test system. The reproducibility's EVM at the fundamental frequency for the broadband testbed considered was found to be within 0.2%.

2.8 Pulsed-RF small- and large-signal measurements

With the trend of increasing power-handling capacity as well as shrinking device dimensions, the generation of self-heating during the measurements of traditional *I–V* characteristics has become an acute problem. In addition, the charging and discharging

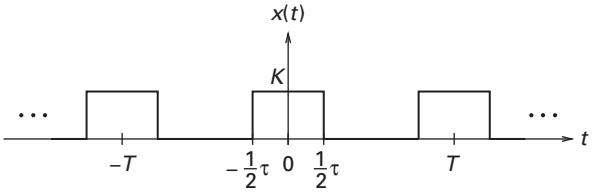


Figure 2.23 Pulse signal with the pulse width τ and pulse period T.

of the traps associated with the device and material imperfections place a common limitation on such DC measurements. To compensate for the slow memory effects caused by self-heating and trapping, pulsed *I–V* measurements provide a very effective approach because they maintain isothermal and iso-trapping conditions during the measurements.

Pulsed-RF measurements are also beneficial for the characterization of power amplifiers developed for use in time-domain multiplexing access communication systems or in radar systems. In addition, such pulsed-RF measurements are valuable for the characterization of devices under higher-bias conditions for which continuous RF wave measurements with constant biasing are not possible due to breakdown in devices.

2.8.1 Analysis of pulsed-RF signals

Figure 2.23 shows the square-wave signal (pulse train) which can be obtained by the convolution of two signals, namely $K\,\mathrm{rect}(t/\tau)$ and $\sum_{n=-\infty}^{n=\infty}\delta(t-nT)$, where τ and T are the width and the period of the pulse signal, respectively. The analytic representation of the signal $x(t)$ is

$$x(t) = \begin{cases} K, & \text{when } -\tau/2 < t - nT < \tau/2, \\ 0, & \text{otherwise,} \end{cases} \tag{2.22}$$

where n is an integer. Since the signal is periodic, the spectrum of the signal can be obtained from a Fourier series analysis. The Fourier series coefficient C_n [25] is found to be

$$C_n = \frac{K\tau}{T}\,\mathrm{sinc}\left(\frac{n\tau}{T}\right), \tag{2.23}$$

where $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. For $K = 1$, $\tau = 1$, and $T = 10$, the magnitude spectrum of C_n is represented in Figure 2.24. The frequency-domain spectrum of the pulse signal is a sampled sinc function with frequency spacing equal to the pulse repetition frequency (PRF), where $\mathrm{PRF} = 1/T$. The distance of the first null ($n = 10$ in Figure 2.24) from the origin in the spectrum ($n = 0$) corresponds to a frequency $1/\tau$. The bandwidth of the main (center) lobe is therefore $2/\tau$.

It is clear in Figure 2.24 that the spectrum of the pulses extends from $n = -\infty$ to $n = \infty$ and this can cause some aliasing error if the LO frequency is not selected appropriately. Note that by definition pulsing a signal leads to a reduction of the *average* signal power by

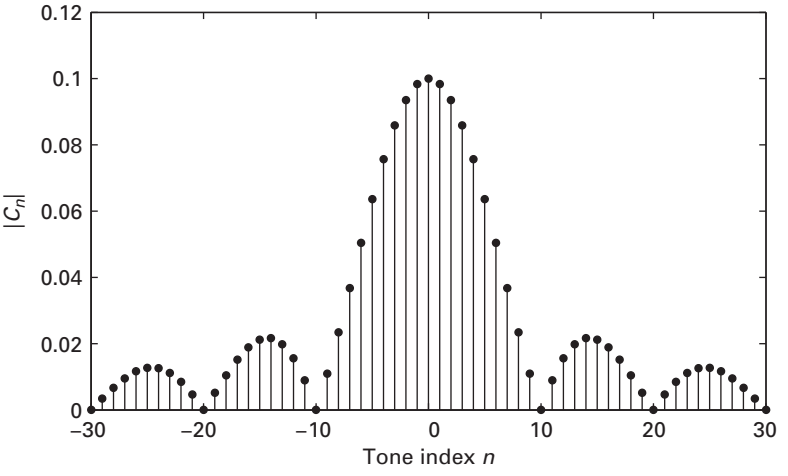


Figure 2.24 Spectrum magnitude for a square-wave signal plotted versus n for $\tau = 1$ and $T = 10$.

$$\text{Power Reduction} = -10\log\left(\frac{\text{Pulse Power}}{\text{CW Power}}\right) = -10\log(\text{Duty Rate}).$$

The total energy in the spectrum is indeed proportional to the duty rate (or duty cycle), which is given as the ratio of the on time of the pulse to the period of the pulse:

$$\text{Duty Rate} = \frac{\tau}{T}. \tag{2.24}$$

This total energy is spread over all the tones of the spectrum. Consequently pulsing the RF signal also yields a decrease of the power of the center tone, which is given by the square of the amplitude of the Fourier series coefficient C_1. The amount of the decrease in the central spectral line caused by the pulsing compared with the CW excitation is therefore

$$\text{Desensitization} = -10\log\left(\frac{|C_1|^2}{K^2}\right) = -20\log(\text{Duty Rate}). \tag{2.25}$$

This is called pulse desensitization [26] [27] [28]. The desensitization is 40.0 dB for 1% duty rate and 49.6 dB for 0.33% duty rate. This is a major concern in the development of traditional pulsed-RF measurement systems using a normal network analyzer. This is due to the fact that the narrow IF receiver of a network analyzer only acquires the center tone in the main lobe, inducing a desensitization that reduces the network analyzer's effective dynamic range.

2.8.2 Pulsed *I–V* pulsed-RF measurement system with the LSNA

In general both V_{GS} and V_{DS} pulsed biases are applied at the gate and the drain of a transistor in pulsed *I–V* systems. Figure 2.25 shows schematically the pulsed *I–V* pulsed-RF measurement system implemented using an LSNA. Port 1 is used for the gate, and port 2 for the drain. A current sensor consisting of a resistor or a transformer is

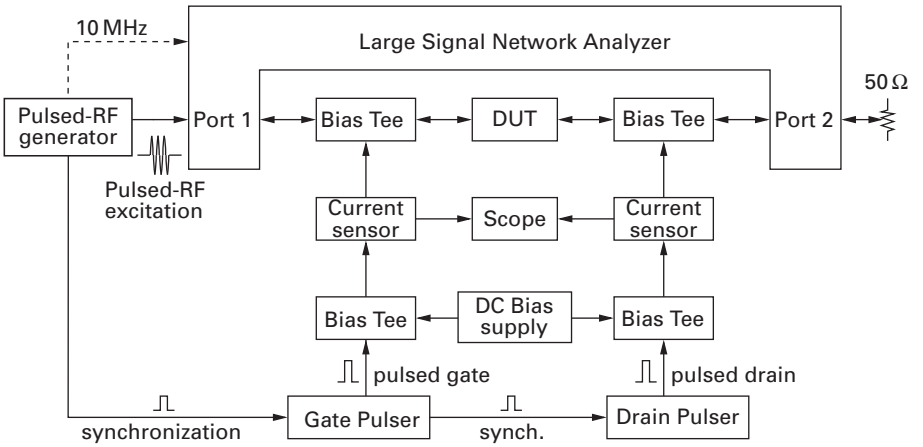


Figure 2.25 Pulsed *I–V* pulsed-RF measurement system with the LSNA.

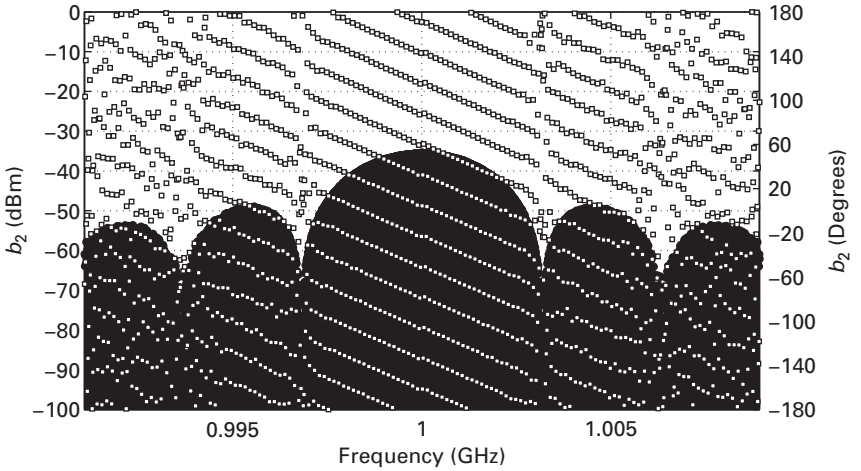


Figure 2.26 Pulsed-RF signal measurements with the LSNA: carrier frequency (f_0) = 1 GHz, modulation frequency (Δf) = 9.918 kHz, and resolution bandwidth 95 Hz. The black circles and the white squares with black outlines represent the magnitude and phase of each tone, respectively.

used to measure the drain current with an oscilloscope. The pulsed-RF signal is provided by a pulsed-RF signal generator. Note that 104 pulses are acquired by the LSNA when using a resolution bandwidth of 95 Hz and modulation frequency 9.918 kHz.

Figure 2.26 shows a generated RF pulse with a duty rate of 0.33% and pulse duration 0.33 μs. The main lobe of the sinc spectrum consists of 601 tones in a bandwidth of about 6 MHz. In addition to generating an RF pulse, the RF signal generator is also triggering the gate and drain pulse generators to overlay the RF pulse upon the pulsed *I–V* biasing signal at the proper moment. The timing for the pulsed *I–V* pulsed-RF signals is shown in Figure 2.27. The drain pulse has a duty rate of 1% and duration 1 μs, which is usually short enough to achieve the targeted isothermal condition.

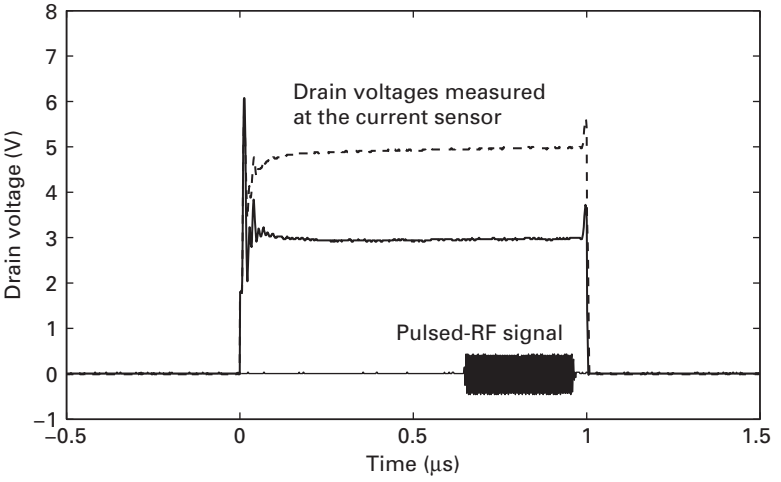


Figure 2.27 Timing for pulsed *I–V* pulsed-RF signals. Pulsed *I–V*: pulse width 1 μs, duty rate 1%. Pulsed RF: pulse width 0.33 μs, duty rate 0.33%.

2.8.3 Measurement bandwidth

To reduce the problem of desensitization in a conventional pulsed-RF system, it would theoretically be desirable to use the entire spectrum of the pulse to obtain higher measurement accuracy. In practice, however, it is not possible to acquire the complete spectrum due to the IF bandwidth limitation of the NVNA used. Nonetheless, by acquiring the spectrum in a wide enough range one can enhance the measurement accuracy.

According to Parseval's theorem, the total average power of the signal $x(t)$ is the sum of the average power in each harmonic component:

$$\frac{1}{T}\int_0^T |x(t)|^2\,dt = \sum_{n=-\infty}^{\infty} |C_n|^2 = C_0^2 + 2\sum_{n=1}^{\infty} |C_n|^2, \tag{2.26}$$

where C_n, the Fourier coefficient for the nth tone, usually has the form of a sinc function if $x(t)$ is a pulsed signal. Thus the power included in the bandwidth range m/τ is

$$\text{Power ratio (\%)} = \frac{\sum_{n=1}^{m} |C_n|^2}{\sum_{n=1}^{\infty} |C_n|^2} \times 100. \tag{2.27}$$

Figure 2.28 shows that the main lobe of a sinc spectrum (m = 300 in this example) includes about 90.3% of the total average power of the sinc spectrum, providing enough power to essentially reduce by half the desensitization experience in the conventional pulsed-RF system compared with the CW case. It is also preferable to focus only on the main lobe because of the noise-floor level of the LSNA. In other words, with a resolution bandwidth of 12 kHz the noise-floor level of the LSNA is typically −70 dB to 20 GHz when the IF frequency is 10 MHz [16], and the deep nulls of the sinc spectrum are usually expected to be below the noise floor as shown in Figure 2.26.

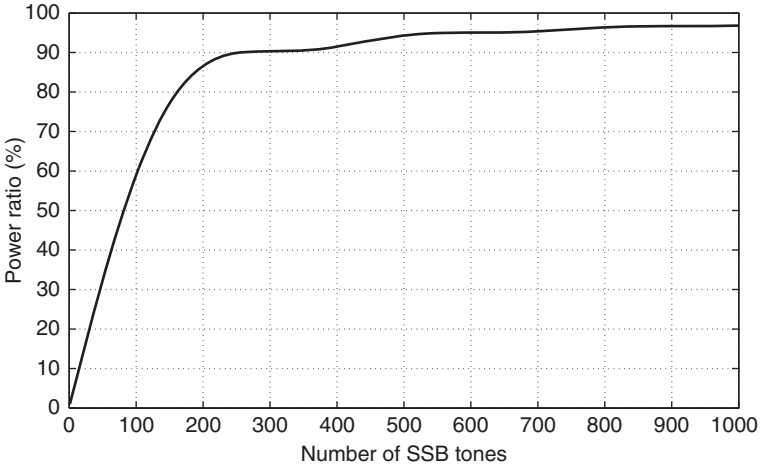


Figure 2.28 Power ratio as a function of single-sideband (SSB) tones. For the analysis, an RF pulse duty rate of 0.33% and a width of 0.33 μs are used.

2.8.4 Envelope analysis of pulsed-RF signals

The NVNA measurements for modulated signals will provide the coefficients $X_i(n\omega_0 + p\omega_\mathrm{m})$ for the waves $x_i(t)$ (the incident $a_i(t)$ or reflected $b_i(t)$ waves) measured at port i:

$$x_i(t) = \mathrm{Re}\left\{\sum_{n=1}^{N}\sum_{p=-\mathrm{SSB}}^{\mathrm{SSB}} X_i(n\omega_0 + p\omega_\mathrm{m})\exp\left[j(n\omega_0 + p\omega_m)t\right]\right\},$$

with N the total number of harmonics acquired and SSB the number of sideband tones acquired. Note that we are excluding the baseband in this expansion, since we are focusing on the RF waves. For pulsed signals it is beneficial to rewrite $x_i(t)$ in terms of the envelope signal $x_i(n\omega_0, t)$ of each of the harmonics:

$$x_i(t) = \mathrm{Re}\left\{\sum_{n=1}^{N} x_i(n\omega_0, t)\exp(jn\omega_0 t)\right\}, \tag{2.28}$$

where the envelope $x_i(n\omega_0, t)$ for the harmonic $n\omega_0$ is defined as

$$\begin{aligned} x_i(n\omega_0, t) &= \sum_{p=-\mathrm{SSB}}^{\mathrm{SSB}} X_i(n\omega_0 + p\omega_\mathrm{m})\exp(jp\omega_\mathrm{m}t) \\ &= E_i^x(n\omega_0, t)\exp\left[j\phi_i^x(n\omega_0, t)\right]. \end{aligned} \tag{2.29}$$

In the above definition $E_i^x(n\omega_0, t)$ and $\phi_i^x(n\omega_0, t)$ are the time-varying amplitude and phase of the envelope waves $x_i(n\omega_0, t)$.

To recover the effective incident and reflected waves when the pulse is on, it is sufficient to select the time t_peak when the incident and reflected waves reach their

peak amplitude. Note that in practice the peaks of the incident and reflected waves are not perfectly synchronized due to the finite propagation group delays τ_{ij} between ports i and j associated with the DUT. It is then normally sufficient to select an average peak time since the DUT group delays are typically much smaller than the pulse duration.

Using this approach the response of the device within the pulse can be obtained for both small- and large-signal excitations. As an example let us consider the evaluation of the S-parameters under small-signal pulsed *I–V* pulsed-RF excitations. By solving the system of equations

$$\begin{aligned} b_1(\omega_0, t_{\text{peak}}) &= S_{11}a_1(\omega_0, t_{\text{peak}}) + S_{12}a_2(\omega_0, t_{\text{peak}}), \\ b_2(\omega_0, t_{\text{peak}}) &= S_{21}a_1(\omega_0, t_{\text{peak}}) + S_{22}a_2(\omega_0, t_{\text{peak}}), \end{aligned} \tag{2.30}$$

for two different measurements with (1) first only port 1 excited by $a_1(\omega_0, t_{\text{peak}})$ and (2) second only port 2 excited by $a_2(\omega_0, t_{\text{peak}})$, we can obtain under small-signal operation the pulsed-RF S-parameters while accounting for non-ideal matching at the non-excited port.

The measured pulsed *I–V* pulsed-RF S-parameters of a GaN HEMT on sapphire are shown in Figure 2.29. In this experiment, a pulsed *I–V* signal (pulse width 1 μs and duty rate 1%) was applied to both the gate and the drain of the transistor. A pulsed-RF signal (pulse width 0.33 μs and duty rate 0.33%) was also injected into either the gate or the drain with a frequency range of 1.0–10.0 GHz in steps of 1 GHz.

Figure 2.30 compares the S-parameters obtained for a GaN HEMT on SiC using a VNA with CW waves (plain lines) and the LSNA with pulsed *I–V* pulsed-RF excitations

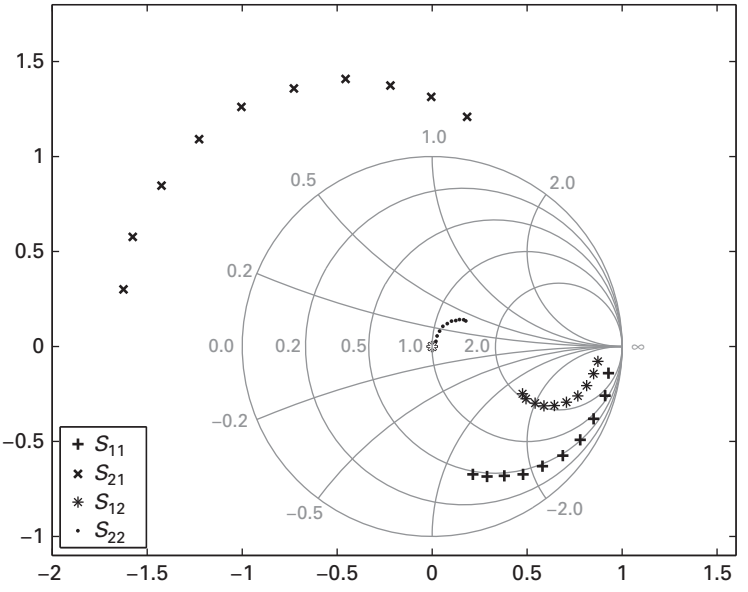


Figure 2.29 Pulsed *I–V* pulsed-RF S-parameters of a GaN HEMT on sapphire (frequency 1–10 GHz, steps of 1 GHz).

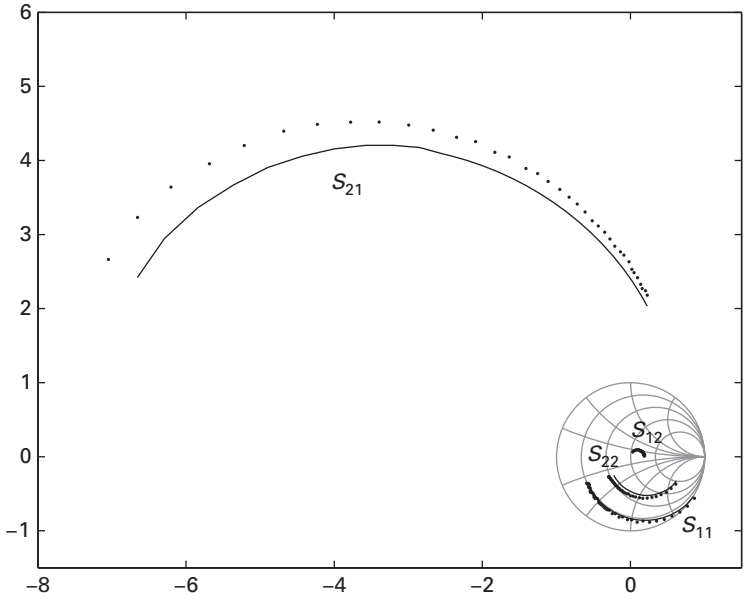


Figure 2.30 Comparison of S-parameters for a GaN HEMT on SiC (frequency 1–10 GHz, steps of 0.25 GHz).

(dots). The frequency ranges from 1.0 to 10.0 GHz with steps of 0.25 GHz. It can be seen that the CW S_{21} measured with the VNA is substantially smaller in amplitude than the S_{21} of the pulsed *I–V* pulsed-RF measurement at low frequencies due to thermal and/or trapping effects.

2.9 Multiple recording of pulsed-RF signals

The broadband measurement method discussed in the previous section for the LSNA, in which all the tones within the main lobe of the signal (90% of the power) are acquired, enables one to partially compensate for the loss of dynamic range [29] by approximately halving the desensitization in dB. However, this method, which requires additional signal processing, including an IF calibration, is still limited in practice to duty cycles above 0.3%. Indeed, as the number of tones increases for decreasing duty cycle, the peak RF power is divided among more tones owing to the lower repetition frequency, and the power of the individual tone will eventually fall below the minimum detectable power level of the LSNA if the peak RF power is kept the same.

Recently, an alternative time-domain technique has been developed and demonstrated at XLIM [30] [31] [32] for the sampler-based NVNA. This allows pulsed-RF measurements with arbitrarily low duty cycles without any reduction in dynamic range. This time-domain technique makes use of the multiple recording feature in modern analog-to-digital converters (ADCs) to control the acquisition of the sample data. A dedicated pulse control board (see Figure 2.32 later) was developed to control the ADC

Table 2.2. Example of parameters used for multiple recording measurement

Parameter	Symbol and value
Fundamental frequency	$f_{RF} = 1\ \text{GHz} = 50 \times f_{LO} + \Delta f_1$
Sampler LO frequency	$f_{LO} = 19.998\ \text{MHz}$
IF frequency	$\Delta f_1 = 100\ \text{kHz} = R_1 f_{RES}$
ADC frequency	$f_{ADC} = 12.8\ \text{MHz}$
FFT size	$N_{FFT} = 16\,384 = 2^{14}$
Resolution frequency	$f_{RES} = 781.25\ \text{Hz} = f_{ADC}/N_{FFT}$
f_1-to-f_{RES} ratio	$R_1 = 128 = 2^7$
IF-to-ADC period ratio	$N_1 = N_{FFT}/R_1 = 16\,384/128 = 128 = 2^7$

acquisition such that the sampled down-converted RF data of the sampler-based NVNA are acquired only at times when the RF pulse is applied [30] [31] [32].

2.9.1 Multiple recording for CW signals

The principal parameters for the multiple-recording time-domain measurement are the ADC frequency f_{ADC}, the local oscillator frequency f_{LO}, and the number of FFT points N_{FFT}. Let us first consider a CW signal consisting of the tone f_{RF} and its harmonics nf_{RF}. The frequencies nf_{RF} are down converted by the sampling down converter to the IF frequencies Δf_n:

$$\Delta f_n = nf_{RF} - S_n f_{LO},$$

with S_n an integer. Let us assume that the IF frequency Δf_n is a multiple R_n of the resolution frequency f_{RES}:

$$\Delta f_n = R_n f_{RES} = nR_1 f_{RES} = n\frac{R_1}{N_{FFT}} f_{ADC},$$

where we assume in this example that the number of harmonics considered is such that we have $n\,\Delta f_1$ smaller than the IF bandwidth. The period for the down-converted IF tone is then

$$T_1 = \frac{1}{\Delta f_1} = \frac{N_{FFT}}{R_1}\frac{1}{f_{ADC}} = \frac{N_1}{f_{ADC}}.$$

We need next to select R_1 such that the ratio $N_1 = N_{FFT}/R_1$ is an integer. Given that N_{FFT} is typically a power of two, this can be achieved with a proper selection of the ADC frequency such that R_1 is also a power of two. An example is shown in Table 2.2.

Using multiple recordings, we can then acquire N_{on} data and wait for p period T_1 (plus N_{on} data) between each acquisition record and the next. As long as N_{on} is an integral fraction R_{on} of N_1, the full IF signal including the fundamental frequency and its harmonics will be acquired. Since N_1 was already selected to be a power of two, it is sufficient to select N_{on} to be a power of two. For example, using $N_{on} = 32 = 2^5$ we acquire the full IF period in $R_{on} = N_1/N_{on} = 4$ records while skipping by p period T_1

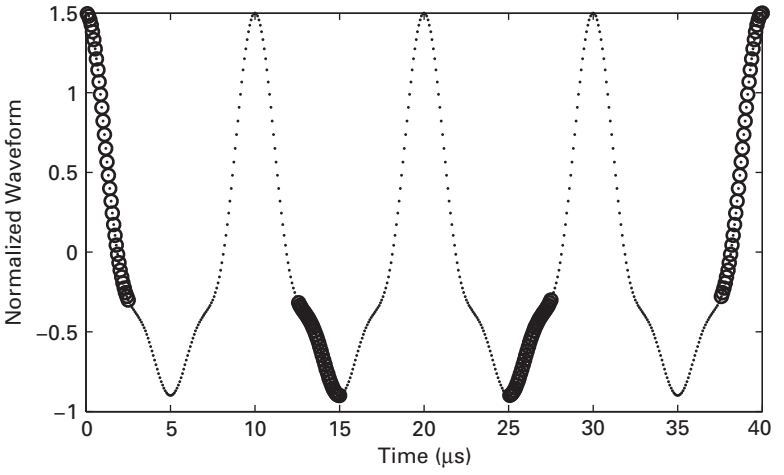


Figure 2.31 Acquisition of the IF tones in a single record (dots) and in four records (circles).

between each recording. This is illustrated in Figure 2.31, where the multiple recorded data are represented by circles. A nonlinear IF signal with three harmonics is rendered in this example to indicate that the harmonics of the RF are also acquired.

The entire operation is fully transparent to the sampler-based NVNA and the same data are acquired for CW excitations using multiple recordings as for the normal acquisition in a single record.

This scheme can now be used to acquire signals that are pulsed (both pulsed RF and pulsed biased). The pulse repetition frequency is then

$$f_{\text{pulse}} = \frac{f_{\text{ADC}}}{p \times N_1 + N_{\text{on}}}.$$

The duty rate can then be set to an arbitrarily small value using a large value of p:

$$\text{Duty Rate} = \frac{N_{\text{on}}}{p \times N_1 + N_{\text{on}}}.$$

Also the data acquired N_{on} can be set to a small value 1 or 2 to sample the RF in a narrow region of the RF *I–V* pulse.

Since the average power of the signal acquired is now the peak power of the pulsed signal, this modified LSNA data acquisition does not theoretically exhibit any degradation in dynamic range for decreasing pulse duty cycle:

$$\text{Desensitization} = 0 \text{ dB}.$$

This 0-dB desensitization is, however, achieved at the price of a longer measurement time and requires additional hardware for the synchronization of the data acquisition.

A schematic diagram for the pulsed *I–V* pulsed-RF measurement setup implemented at OSU with an LSNA using the pulse control board developed at XLIM is shown in Figure 2.32. The DC bias and the pulsed bias are combined with a baseband bias tee. The resulting time-varying bias voltages are then added to the pulsed-RF signal

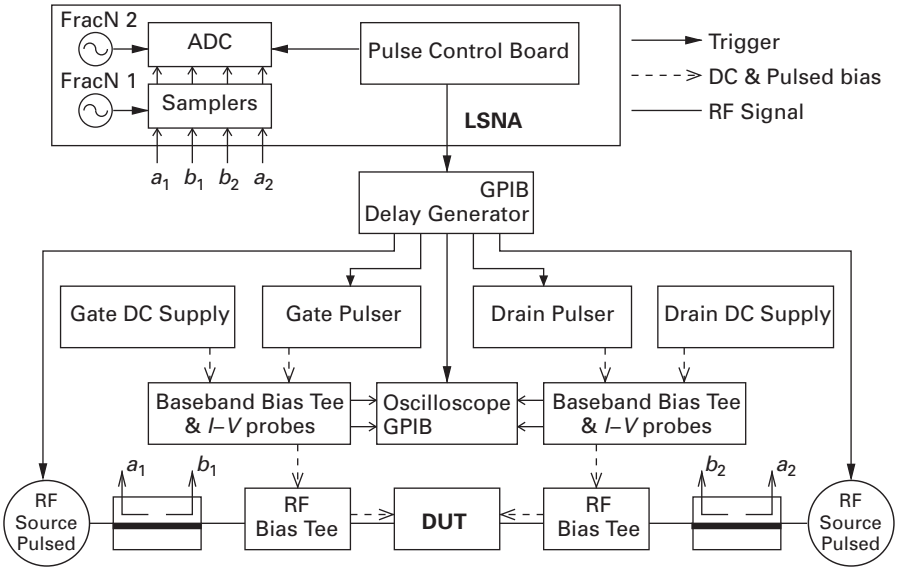


Figure 2.32 Pulsed *I–V* pulsed-RF measurement system with multiple recording using a modified LSNA. (From [33] with permission, ©2010 IEEE.)

generated by the RF sources using an RF bias tee. The combined DC, pulsed, and RF signal is then fed to the device. The synchronization of the pulsed bias and the pulsed RF signal is precisely controlled by the delay generator, which can output separate triggers for different instruments. A trigger is also sent to the oscilloscope for monitoring the timing of the applied pulsed bias and pulsed RF signal. The reference timing for all the triggers output from the delay generator is controlled by an external trigger sent by the pulse control board. At the same time, the pulse control board sends another trigger to the ADC to control the data acquisition. The pulse control board itself includes a tunable delay for the trigger sent to the ADC to ensure that the peak of the pulsed RF signal is measured at the desired time while compensating for the sampler latency. In addition, the pulse control board, the fracN controlling the samplers, and the fracN controlling the ADCs, the oscilloscope, and the delay generator are all locked to the 10-MHz reference clock signal generated by the RF source so that all instruments share the same time base. Note that the clock for the ADC is set by a fracN source that is separate from the one controlling the LSNA samplers so that the sampling and ADC frequencies, locked as well on the same 10 MHz, can be set independently.

Beside synchronizing the data acquisition of the LSNA with the pulsed *I–V* bias applied, the function of the pulse control board developed by XLIM includes making sure that the suspended data acquisition is resumed at the correct sampling time to correctly capture the sampling down-converted RF signal without introducing any spectral leakage [30]. To verify the functionality of the setup and to select the correct data-acquisition time of the ADC, the measured incident and reflected waves of a calibration standard "Through" are compared in Figure 2.33. The peak output power levels of the

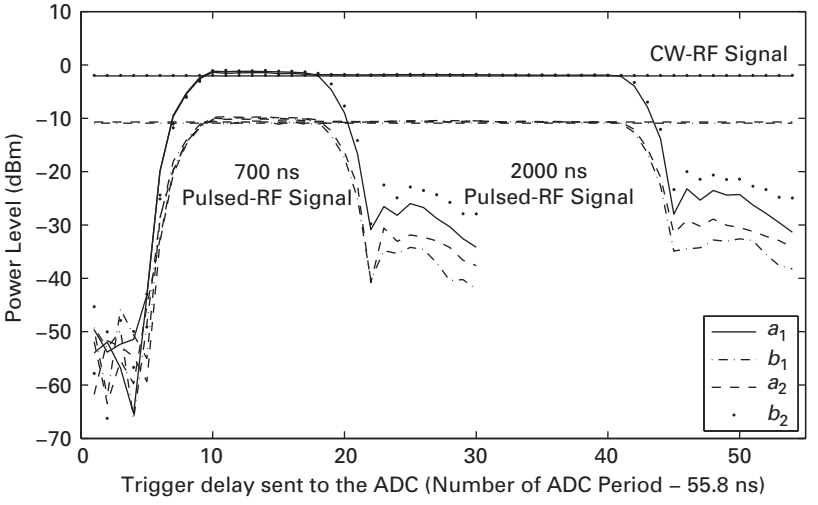


Figure 2.33 Measured traveling waves of the pulsed-RF signal with pulse widths of 700 ns and 2000 ns are compared with those obtained for a CW-RF signal of the same peak power. (From [33] with permission, ©2010 IEEE.)

pulsed-RF signal, with two different pulse widths and the same pulse repetition period, and the CW-RF signal from the RF source are both 0 dBm. It is demonstrated that, with the change of pulse duty cycle, no significant change in the measured peak power level and noise floor is observed. This demonstrates that the setup does not exhibit any degradation in its dynamic range when the pulse duty cycle decreases. Note that the latency in the LSNA is 550 ns. Much smaller latencies are achievable with more recent implementations of the sampler-based NVNA such as the SWAP.

2.9.2 Multiple recording for jointly pulsed and modulated signals

The multiple recording developed at XLIM is a very elegant and powerful technique for acquiring pulsed RF data without any desensitization. The method can be further applied to signals that are both pulsed and modulated (multitone excitations). When performing the measurement of multitone excitations, additional considerations are required in order to handle the modulation [34].

The principle of operation for the multiple recording measurement of modulated signals is as follows. First, the modulation frequency f_{m} needs to be a multiple R_{m} of the resolution frequency f_{RES}, and the modulation period $1/f_{\mathrm{m}}$ needs to consist of an integral number N_{m} of sampling points. This is easily achieved by using a power of two for R_{m} since N_{FFT} is already a power of two. Second, the number of points N_{on} to be sampled when the pulse is on should be selected such that N_{m} is a multiple R_{on} of N_{on}:

$$N_{\mathrm{m}} = \frac{f_{\mathrm{ADC}}}{f_{\mathrm{m}}} = \frac{N_{\mathrm{FFT}}}{R_{\mathrm{m}}} = R_{\mathrm{on}} \times N_{\mathrm{on}}.$$

This is again easily achieved by selecting a power of two for N_{on} since N_m was selected to be a power of two. The pulse repetition period can then be set to any integer multiple of p of the modulation period $1/f_m$ plus N_{on} such that the required pulse repetition frequency can be set to any value given by

$$f_{pulse} = \frac{f_{ADC}}{p \times N_m + N_{on}},$$

with p an integer. For example, consider a modulation frequency of 12.5 kHz for an RF signal with a fundamental frequency of 1 GHz. Using the LSNA settings given in Table 2.2, we have $R_m = 2^4$ and $N_m = N_{FFT}/R_m = 2^{10}$, which are both integers, as required. We can then, for example, select $N_{on} = 2^4$ for the number of points to be sampled in each record, since this yields an integral number $R_{on} = N_m/N_{on} = 2^6$ of multiple records needed to acquire a complete period $1/f_m$ of the modulated signal while skipping p period $1/f_m$.

Again it has been verified experimentally [30] that this modified multiple-recording acquisition technique correctly captures the sampling down-converted RF signals without introducing any spectral leakage even though the RF signals are both jointly pulsed and modulated. The phase repeatability is within a fraction of a degree. This measurement technique will be applied in Section 5.6 to extend the real-time active load-pull (RTALP) measurement technique to pulse operation. The resulting pulsed-RF RTALP scheme permits one to rapidly generate P1dB contour plots for transistors excited by pulsed-RF signals. Intermodulation tones and spectral regrowth can also be characterized for jointly pulsed and modulated RF signals using this extended multiple-recording technique.

References

[1] P. D. Hale, C. M. Wang, D. F. Williams, K. A. Remley, and J. Wepman, "Compensation of random and systematic timing errors in sampling oscilloscopes," *IEEE Transactions on Instrumentation and Measurements*, Vol. 55, No. 6, pp. 2146–2154, Dec. 2006.

[2] T. Williams, O. Mojon, S. Woodington, L. Simon, J. Lees, M. Barciela, F. J. Benedikt, and P. J. Tasker, "A robust approach for comparison and validation of large signal measurement systems," in 2008 IEEE MTT-S International *Microwave Symposium Digest*, pp. 257–260, June 2008.

[3] D. Barataud, M. Campovecchio, and J.-M. Nebus, "Optimum design of very high-efficiency microwave power amplifiers based on time-domain harmonic load-pull measurements," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 49, No. 6, 1107–1112, June 2001.

[4] P. Blockley, D. Gunyan, and J. B. Scott, "Mixer-based, vector-corrected, vector signal/network analyzer offering 300 kHz–20 GHz bandwidth and traceable phase response," in *Digest, 2005 IEEE MTT-S International Microwave Symposium*, CDROM, 2005.

[5] Network Measurement and Description Group: www.nmdg.be.

[6] Agilent Technologies: www.agilent.com.

[7] J. Verspech, P. Debie, A. Barel, and L. Martens, "Accurate on wafer measurements of phase and amplitude of the spectral components of incident and scattered voltage waves at the signal ports of a nonlinear microwave device," in *1995 IEEE MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1029–1032, 1995.

[8] J. Verspecht, "Large-signal network analysis," *IEEE Microwave Magazine*, Vol. 6, No. 4, pp. 82–92, Dec. 2005.

[9] Verspect–Teyssier–DeGroote s.a.s: www.vtd-rf.com.

[10] S. K. Myoung, D. Chaillot, P. Roblin, W. Dai, and S. J. Doo, "Volterra characterization and predistortion linearization of multi-carrier power amplifiers," in *64th ARFTG Conference Digest*, 2004.

[11] J. Verspecht, "Calibration of a Measurement System for High Frequency Nonlinear Devices," Doctoral Dissertation, Vrije Universiteit Brussel, 1995.

[12] J. Verspecht "Broadband sampling oscilloscope characterization with the 'nose-to-nose' calibration procedure: a theoretical and practical analysis," *IEEE Transactions on Instrumentation and Measurement*, Vol. 44, No. 6, pp. 991–997, Dec. 1995.

[13] V. Teppati and A. Ferrero, "A new class of nonuniform, broadband, nonsymmetrical rectangular coaxial-to-microstrip directional couplers for high power applications," *IEEE Microwave Wireless Component Letters*, Vol. 13, No. 4, pp. 152–154, 2003.

[14] F. De Groote, J. Verspecht, C. Tsironis, D. Barataud, and J.-P. Teyssier, "An improved coupling method for time domain load-pull measurements," in *65th ARFTG Conference Record*, pp. 57–60, June 2005.

[15] S. J. Doo, "New Pulsed-*IV* Pulsed-RF Measurement Techniques for Characterizing Power FETs for Pulsed-RF Power Amplifier Design," Dissertation, Doctor of Philosophy, Ohio State University, 2008.

[16] NMDG Engineering, *Large-Signal Network Analyzer System Manual*, 2004.

[17] T. Reveyrand, A. Mallet, F. Gizard, L. Lapierre, J.-M. Nébus, and M. Vanden Bossche, "A new time domain waveform measurement setup to investigate internal node voltages in MMICs," in *Microwave Technology and Technique Workshop*, ESTEC, Noordwijk, 2004.

[18] G. Gonzalez, *Microwave Transistor Amplifiers – Analysis and Design*, Prentice Hall, 1997.

[19] J. Verspecht, Presentation on the uncalibrated acquisition with an LSNA of signals with several gigahertz modulation bandwidth, Nonlinear Measurement Workshop at ARFTG 66th Conference in December 2005, Washington, DC.

[20] W. Van Moer and Y. Rolain, "An improved broadband conversion scheme for the large-signal network analyzer," *IEEE Transactions on Instrumentation and Measurement*, Vol. 58, No. 2, pp. 483–487, Feb. 2009.

[21] G. Avolio, G. Pailloncy, D. Schreurs, M. Vanden Bossche, and B. Nauwelaers, "On-wafer LSNA measurements including dynamic-bias," in *Proceedings of the 39th European Microwave Conference*, 2009.

[22] K. Remley, D. Schreurs, D. Williams, and J. Wood, "Extended NVNA bandwidth for long-term memory measurement," in *2004 IEEE MTT-S Digest*.

[23] Y. Ko, P. Roblin, S. Myoung, J. Strahler, F. De Groote, and J. P. Teyssier, "Multiharmonic broadband measurements using a large signal network analyzer," in *75th ARFTG Conference Digest*, 2010.

[24] M. El Yaagoubi, G. Neveux, D. Barataud, T. Reveyrand, J.-M. Nebus, F. Verbeyst, F. Gizard, and J. Puech, "Time-domain calibrated measurements of wideband multisines using a large-signal network analyzer," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 56, No. 5, pp. 1180–1192, May 2008.

[25] S. S. Soliman and M. D. Srinath, *Continuous and Discrete Signals and Systems*, Prentice Hall, 1990.

[26] Agilent Technologies, *Pulsed Measurements with the Agilent 8720ES and 8753ES Network Analyzer – Product Note*, 2000.

[27] Agilent Technologies, *Using a Network Analyzer to Characterize High-Power Components – Application Note 1287-6*, 2003.

[28] Agilent Technologies, *PNA Microwave Network Analysis – Application Note 1408-11*, 2004.

[29] S. J. Doo, P. Roblin, S. Lee, D. Chaillot, and M. Vanden Bossche, "Pulsed-IV pulsed-RF measurements using a large signal network analyzer," in *65th ARFTG Conference*, June 2005.

[30] F. De Groote, O. Jardel, T. Reveyrand, J.-P. Teyssier, and R. Quéré, "Very small duty cycles for pulsed time domain transistor characterization," *Proceedings of the European Microwave Association*, Vol. 4, No. 6, pp. 112–117, June 2008.

[31] J. P. Teyssier, C. Charbonniaud, D. Barataud, M. Nébus, and R. Quéré, "Large-signal time-domain characterization of microwave transistors under RF pulsed conditions," 2nd Nonlinear Measurement Workshop, 60th ARFTG Conference in December 2002, Washington, DC.

[32] J. Faraj, F. De Groote, J.-P. Teyssier, J. Verspecht, R. Quéré, and R. Aubry, "Pulse profiling for AlGaN/GaN HEMTs large signal characterizations," *38th European Microwave Conference EuMC 2008*, pp. 757–760, 2008.

[33] C. K. Yang, P. Roblin, F. D. Groote, J. P. Teyssier, S. Ringel, C. Poblenz, Y. Pei, J. Speck, and U. Mishra, "Pulsed-*IV* pulsed-RF cold-FET parasitic extraction of biased AlGaN/GaN HEMTs using large signal network analyzer," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 58, No. 5-I, pp. 1077–1088, May 2010.

[34] F. De Groote, P. Roblin, J. P. Teyssier, C. Yang, S. Doo, M. Vanden Bossche, "Pulsed multitone measurements for time domain load pull characterizations of power transistors," in *73rd ARFTG Conference Digest*, Boston, 2009.

3 Device modeling and verification with NVNA measurements[1]

Compact models of active devices that are integrated into circuit simulators are an integral part of the design process of both RF integrated circuits (RFICs) and monolithic microwave integrated circuits (MMICs). The accuracy of these device models in capturing the range of behavior of the devices is of critical importance for the successful design of first-pass RFICs and MMICs. Reliable compact device models have also come to be expected by microwave engineers for the design of RF/microwave circuits using discrete components such as power amplifiers, mixers, and oscillators.

This chapter will present some examples of the applications of NVNA measurements to both device modeling and model verification. The focus will be placed on devices for which memory effects are not dominant. The characterization and modeling of memory effects will be discussed in the next chapter.

NVNAs are primarily used for model verification, and an example will be presented. Various RF excitations can be used to test the devices in normal mode of operation, in breakdown [1] [2], or to investigate the device symmetry or lack thereof. Finally, using even more exotic modulated excitations, NVNA measurements can permit one to rapidly capture, in a single measurement, the operation of the device in a wide range of voltages for the direct extraction of nonlinear models.

3.1 Model verification

NVNAs can be beneficially used for validating compact device models for large-signal excitations [3]. Compact device models are usually extracted using *I–V* (DC or pulsed) characteristics and small-signal (CW or pulsed-RF) S-parameters for a wide range of biasing conditions (V_{GS} and V_{DS} for three-terminal FETs) and temperatures. The general validity of the model topology and the accuracy of the voltage dependences of the current sources and charges extracted need to be validated with large-signal measurements.

Before the availability of NVNAs, mostly scalar figures of merit were used, including harmonic generation, intermodulation distortion (IMD), and, in the most exacting case, load-pull data. With the coming of NVNAs, both the amplitude and the phase of the

[1] Research collaboration with Seok Joo Doo and Venkatesh Balasubramanian is gratefully acknowledged.

harmonics and IMD are now available for model validation. Also the actual measured current and voltage waveforms and the dynamic loadlines at RF can now be compared with the simulation results.

Designers of high-speed integrated circuits are typically provided with several models that are representative of the statistical variation of the device performance to be expected from die to die. Typically models for weak, medium, and strong devices are used to facilitate the evaluation of the performance of the circuits designed using the so-called "corner" analysis and to optimize the yield using design centering.

To illustrate possible model verifications for integrated-circuit design, we consider the case of a 65-nm CMOS process for which BSIM4 is used for the device modeling [4]. The device selected has a gate length of 70 nm, gate (finger) width of 1 μm and 16 fingers [5]. We shall present results for a common-source device operating in classes A and B. Details on the LSNA testbed used for the class-A and class-B measurements will be given in Chapter 5. Common gate results will be presented in the next section. For the class-A verifications, operation in both the saturated and linear regions will be considered. The focus will be placed on the current and voltage waveforms and associated dynamic loadlines.

The LSNA measurements were conducted at 4 GHz with four harmonics recorded, corresponding to 4, 8, 12, and 16 GHz. The current and voltage waveforms are reconstructed using the amplitude and phase of these four tones. The DC *I–V* characteristic for the device measured is shown in Figure 3.1, where it is compared with those of the weak and normal foundry models.

The harmonic balance simulations were performed using 20 harmonics (order 20). It was verified that a satisfactory convergence occurred with 15 harmonics. However, when comparing the measured and simulated waveforms, only the first four harmonics

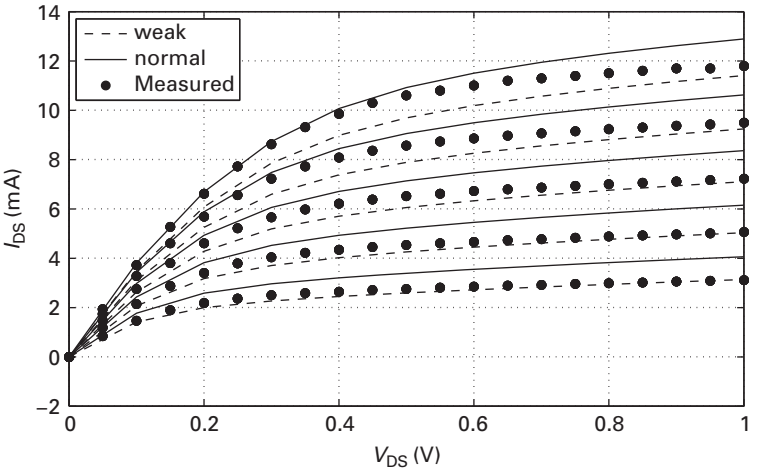


Figure 3.1 Comparison of measured (dots) and modeled (dashed line and plain line) DC *I–V* characteristics for a 70-nm MOSFET. V_{GS} is varied from 0 to 1.2 V in steps of 0.3 V.

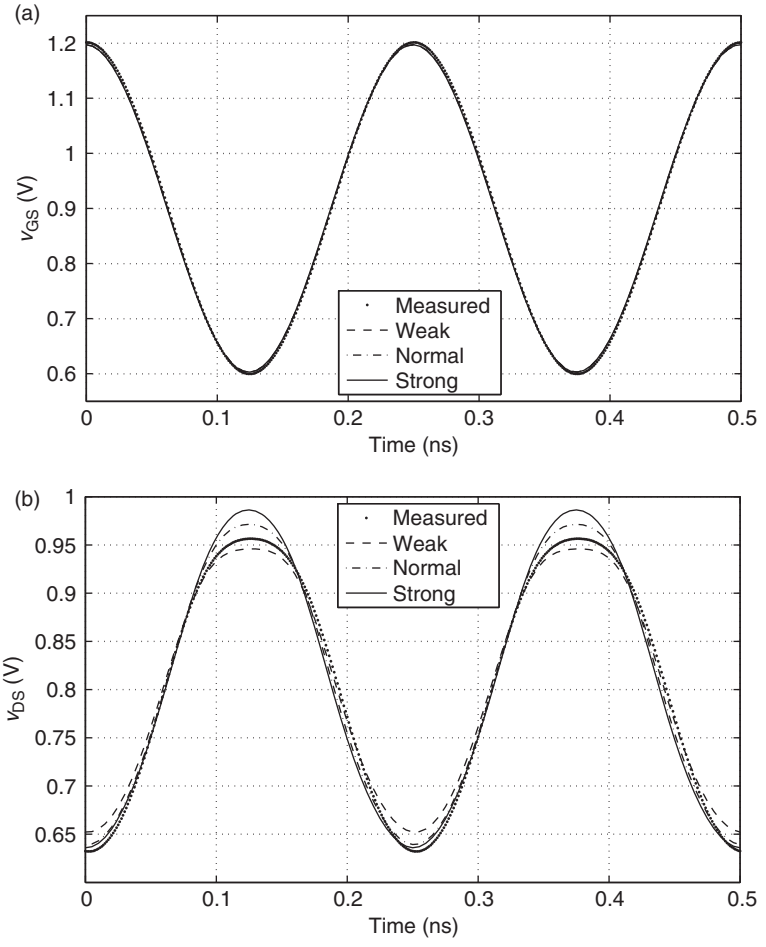


Figure 3.2 Class-A amplifier operating in the saturation region. (a) Input voltage waveform: v_{GS} versus time. (b) Output voltage waveform: v_{DS} versus time.

(calculated using simulations of order 20) were used for reconstructing the waveforms in order to obtain a more meaningful comparison.

Figure 3.2 shows the gate-to-source voltage waveform $v_{GS}(t)$ and drain-to-source voltage waveform $v_{DS}(t)$ for class-A operation in the saturation region. Note that, although a sinusoidal incident wave $a_1(t)$ is applied, nonlinearities may appear in the reflected waves $b_1(t)$ and $b_2(t)$ and non-sinusoidal voltage and current waveforms will result. A small voltage saturation is noticeable in the peak of $v_{DS}(t)$ in the measured data (dotted line). Figure 3.3(a) shows the drain-current waveform $i_D(t)$. A small current saturation is noticeable at low values of the drain current $i_D(t)$ due to the device approaching pinchoff. Figure 3.3(b) shows the dynamic transfer characteristic: $i_D(t)$ versus $v_{GS}(t)$. The nonlinearity of the transconductance is observable. Figure 3.3(c) shows the dynamic output loadline: $i_D(t)$ versus $v_{DS}(t)$. A linear trajectory is obtained because a resistive load (50 Ω) is used. The comparison with the three corner models

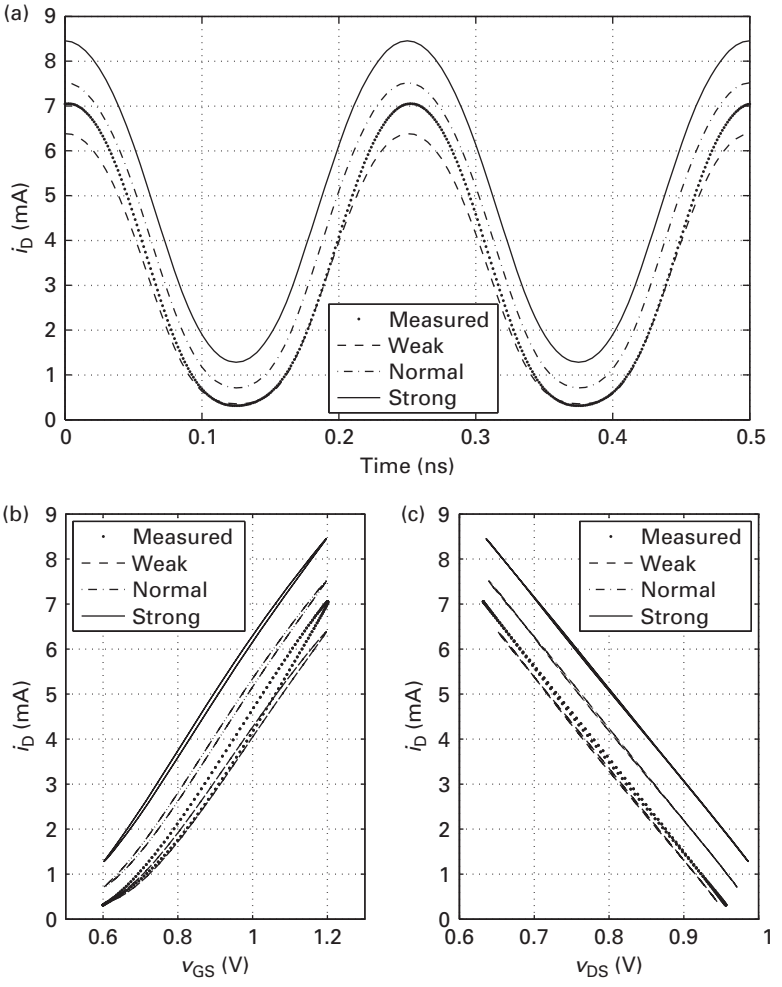


Figure 3.3 Class-A amplifier operating in the saturation region: (a) current waveform $i_D(t)$ versus time, (b) dynamic transfer characteristic i_D versus v_{GS}, and (c) dynamic output loadlines i_D versus v_{DS}.

indicates that the device performance is mostly between that of the weak (dashed line) and the normal (dash–dotted line) models for the particular device measured.

Figure 3.4 shows the gate-to-source voltage waveform $v_{GS}(t)$ and the drain-to-source voltage waveform $v_{DS}(t)$ for class-A operation in the linear region. Figure 3.5(a) shows the drain-current waveform $i_D(t)$. Despite the sinusoidal input voltage, a small saturation is detected at both low and high values of $v_{DS}(t)$ and $i_D(t)$ in the measured data (dotted line). This saturation is due to the device approaching pinchoff at low gate voltages and to the nonlinearity of the transconductance in the linear region at low drain voltages and high gate voltages. Figure 3.5(b) shows the dynamic transfer characteristic: $i_D(t)$ versus $v_{GS}(t)$. The nonlinearity of the transconductance is observable

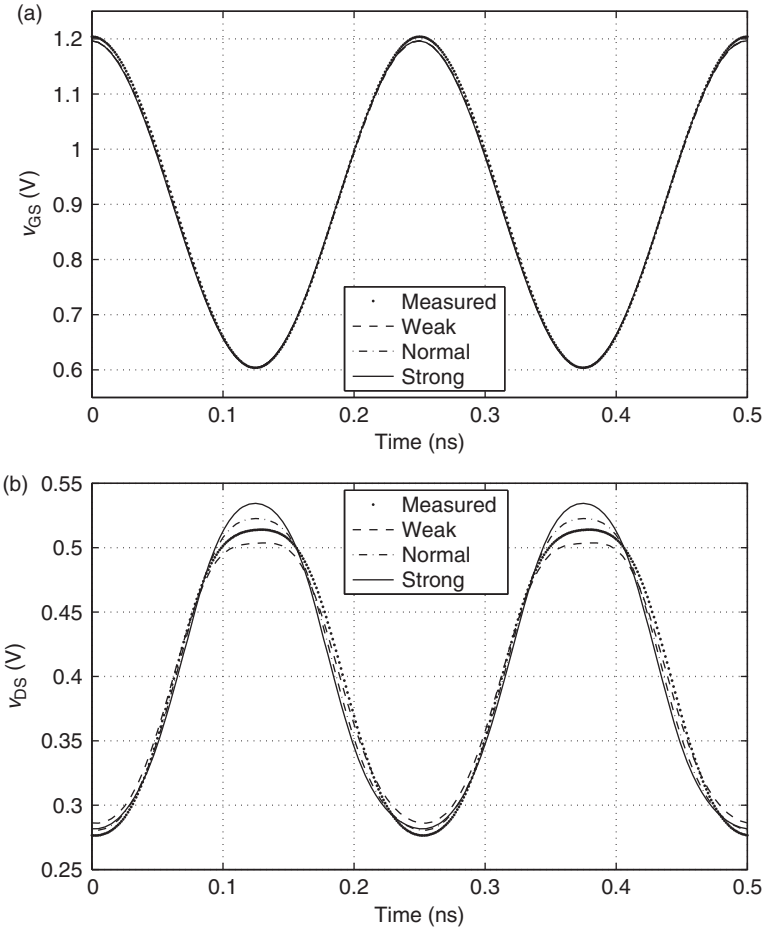


Figure 3.4 Class-A amplifier operating in the linear region. (a) Input voltage waveform: v_{GS} versus time. (b) Output voltage waveform: v_{DS} versus time.

at both low and high values of the gate voltage. Figure 3.5(c) shows that a linear trajectory is again obtained for the dynamic drain loadline $i_D(t)$ versus $v_{DS}(t)$ due to the resistive load at the drain terminal. The comparison with the three corner models indicates that the measured device is again between the weak (dashed line) and the normal (dash–dotted line) models. However, the drain current is seen to reach slightly below zero current, whereas the BSIM4 models remain clearly above zero current. This may arise due to memory effects associated with the charge storage in the drain-to-source capacitance.

Figures 3.6 and 3.7 report on near-class-B operation in the saturation region. Class-B operation has been approximated by shorting the second harmonic. As we shall see in Chapter 5, this can be accurately realized using an active load-pull testbed. Higher harmonics could also be shorted, but this is usually less critical. Figure 3.6 shows the gate-to-source voltage waveform $v_{GS}(t)$ and the drain-to-source voltage waveform

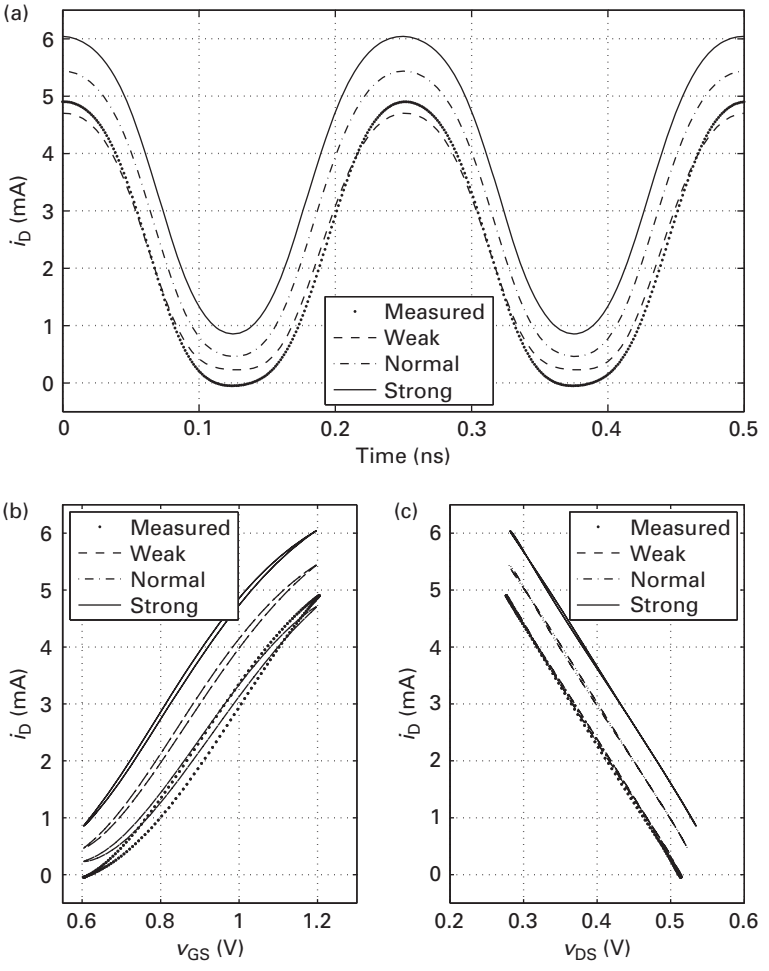


Figure 3.5 Class-A amplifier operating in the linear region: (a) current waveform i_D versus time, (b) dynamic transfer characteristic i_D versus v_{GS}, and (c) dynamic output loadlines i_D versus v_{DS}.

$v_{DS}(t)$. Figure 3.7(a) shows the drain current waveform $i_D(t)$. The device is clearly off for half of its cycle. Some of the ringing in the current toward zero is due to the limitation to four harmonics. Figure 3.7(b) shows the transfer characteristic: $i_D(t)$ versus $v_{GS}(t)$. Figure 3.7(c) shows the dynamic output loadline: $i_D(t)$ versus $v_{DS}(t)$. The dynamic loadline is no longer linear due to the device turning off during half of the RF cycle. The comparison with the three corner models indicates that the measured device is operating closer to the normal (dash–dotted line) model.

The various measurements presented above for class-A and class-B operations give an example of the verification of a foundry model using waveforms and dynamic loadlines. More verification tests for device symmetry and MOSFETs with varying gate width will be presented in the next sections.

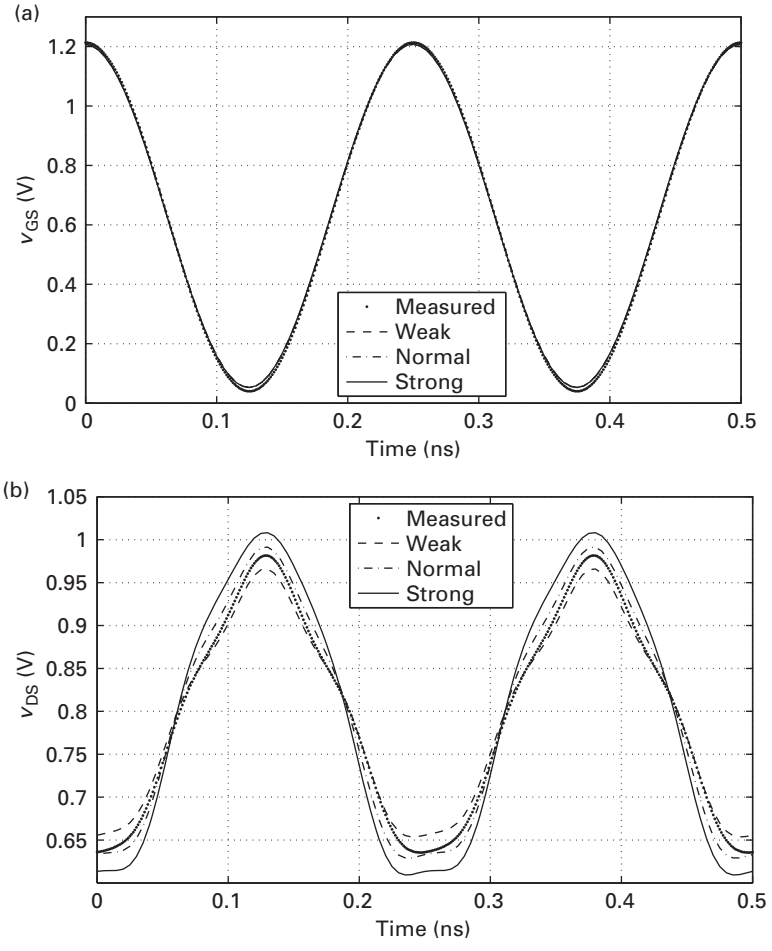


Figure 3.6 Class-B amplifier operating in the saturation region. (a) Input voltage waveform: v_{GS} versus time. (b) Output voltage waveform: v_{DS} versus time. (From [6] with permission, ©2009 IEEE.)

3.2 Model symmetry

Model symmetry is of importance for a number of circuits such as resistive mixers. To discuss the model symmetry, we shall first introduce a charge-based model in the v_{GS} and v_{DS} representation. In the conventional charge-based model for a three-terminal FET, the total gate, drain, and source currents are given by

$$\begin{aligned} i_G &= i_{disp,G}, \\ i_D &= I_D(v_{GS}, v_{DS}) + i_{disp,D}, \\ i_S &= i_D + i_G, \end{aligned}$$

where I_D is the DC drain current, and $i_{disp,G}$ and $i_{disp,D}$ are the gate and drain displacement currents, respectively. These displacement currents $i_{disp,X}$ are associated with the charge element Q_X (Q_G or Q_D) and derived from them using

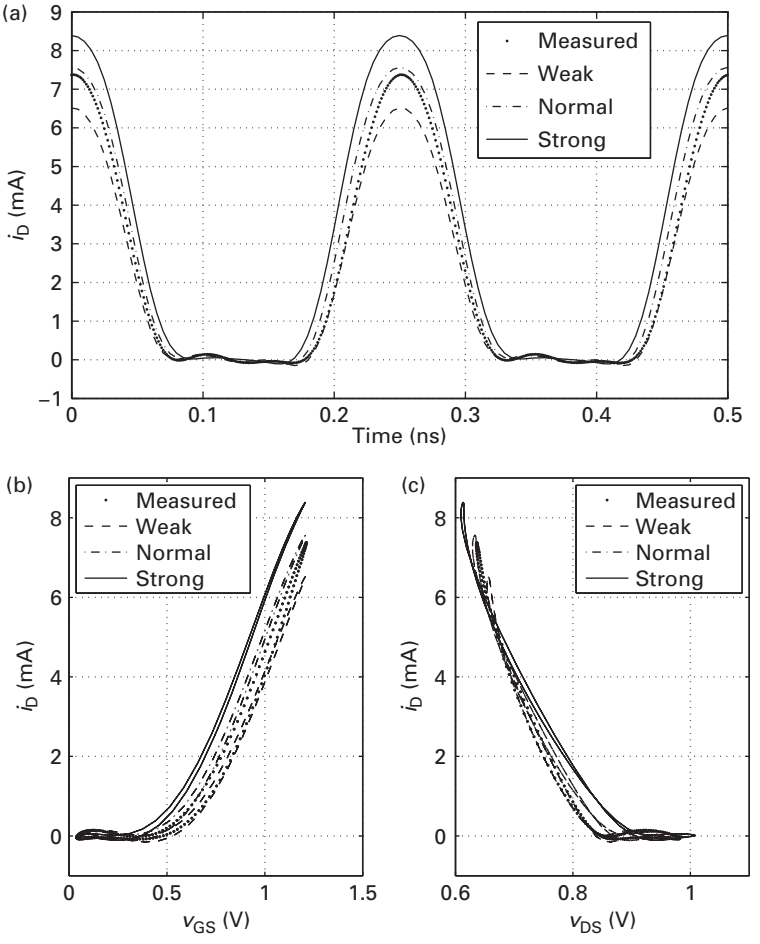


Figure 3.7 Class-B amplifier operating in the saturation region: (a) current waveform i_D versus time, (b) dynamic transfer characteristic i_D versus v_{GS}, and (c) dynamic output loadline i_D versus v_{DS}. (From [6] with permission, ©2009 IEEE.)

$$i_{\text{disp,X}} = \frac{dQ_X(v_{GS}, v_{DS})}{dt} - \frac{d}{dt}\left[\tau_X(v_{GS}, v_{DS})\, i_{\text{disp,X}}\right].$$

Note that τ_X is the non-quasi-static charge-redistribution time-constant associated with the charge-element Q_X.

Figure 3.8(a) shows a circuit topology for implementing this charge-based large-signal model which relies on two non-quasi-static time-constants (τ_G and τ_D). An alternative and equivalent topology is also shown in Figure 3.8(b). In this alternative topology, the current I' and charges Q' are given by

$$I'_D(v_{GS}, v_{GD}) = I_D(v_{GS}, v_{DS}),$$
$$Q'_D(v_{GS}, v_{GD}) = Q_D(v_{GS}, v_{DS}), \tag{3.1}$$
$$Q'_S(v_{GS}, v_{GD}) = -Q_G(v_{GS}, v_{DS}) - Q_D(v_{GS}, v_{DS}). \tag{3.2}$$

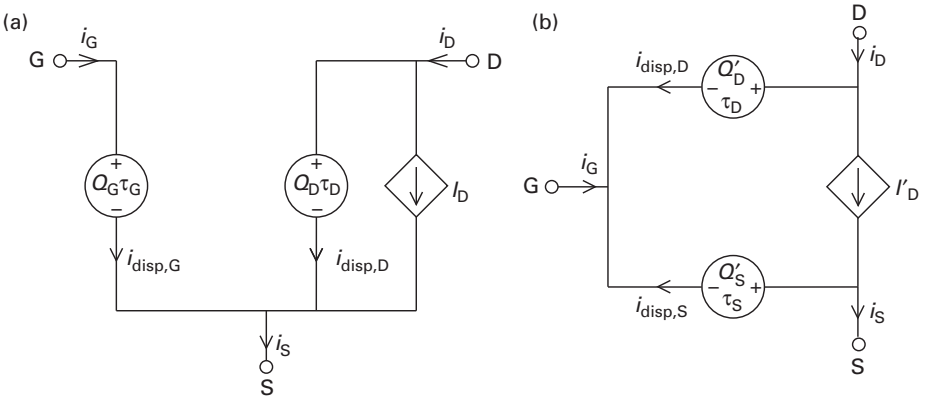


Figure 3.8 Charge-based large-signal model for the intrinsic FET with gate and drain charges (a) and with gate and source charges (b).

In the limit of $\tau_G = \tau_D = \tau_S = \tau$ these two topologies are equivalent.

Usually the FET model is extracted for positive external drain-to-source voltages ($v_{DS} > 0$) and the model is extended to negative external drain voltages for a symmetric device by switching the internal source and drain terminals of the model. It results that the internal v'_{GS} and v'_{DS} are remapped as follows when switching from positive to negative external v_{DS} voltages:

$$\begin{aligned} &\text{Positive } v_{DS} \rightarrow \text{Negative } v_{DS},\\ &v'_{GS} = v_{GS} \rightarrow v'_{GS} = v_{GD} = v_{GS} - v_{DS},\\ &v'_{DS} = v_{DS} \rightarrow v'_{DS} = -v_{DS}. \end{aligned}$$

Alternatively, if no switch is used and the current and charge functional representations of the model are to handle negative drain voltages, then the symmetry assumption places some constraints on these voltage dependences when expressed in terms of the external voltages [7]:

$$\begin{aligned} I_D(v_{GS}, v_{DS}) &= -I_D(v_{GS} - v_{DS}, -v_{DS}),\\ Q_G(v_{GS}, v_{DS}) &= Q_G(v_{GS} - v_{DS}, -v_{DS}),\\ Q_D(v_{GS}, v_{DS}) &= -Q_D(v_{GS} - v_{DS}, -v_{DS}) - Q_G(v_{GS} - v_{DS}, -v_{DS}). \end{aligned} \tag{3.3}$$

Note that, following Ref. [8], a simpler extraction and mathematical representation will result if we switch from common-source state-variables (v_{GS}, v_{DS}) to common-gate state-variables (v_{GS}, v_{GD}) and use the charges Q'_D and Q'_S defined in Equations (3.1) and (3.2). Further, this representation facilitates the characterization and modeling of non-symmetric devices. Note that the non-quasi-static topology shown in Figure 3.8(a) using the non-quasi-static time-constants τ_G and τ_D can still be used, since the charges Q_D and Q_S can be expressed in terms of the modeled charges Q'_D and Q'_S.

NVNA measurements with a common-gate FET provide a methodology for verifying the validity of the assumption of symmetry. The common-gate FET topology is shown

Table 3.1. Possible amplitudes and phases for the applied excitations at ports 1 and 2

Case	V_1	V_2
1	$\lvert V_1\rvert \angle 0°$	$\lvert V_1\rvert \angle 0°$
2	$\lvert V_1\rvert \angle 0°$	$\lvert V_1\rvert \angle 180°$
3	$\lvert V_1\rvert \angle 0°$	$\lvert V_1\rvert \angle 90°$

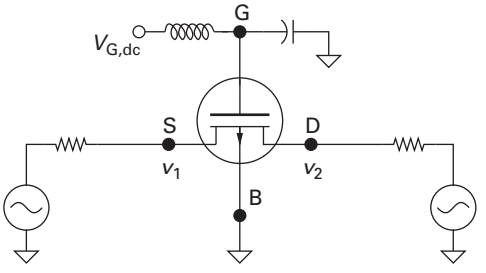


Figure 3.9 Common-gate FET for symmetry testing.

in Figure 3.9. An RF ground is established at the gate terminal using a capacitance to ground. This RF grounding still permits one to change the DC gate voltage. The body terminal is also grounded. Two RF sources are connected at the source and drain input terminals. They are both phase-locked to the NVNA via the 10-MHz reference channel. The source and drain excitations are set to have the same amplitude but with different phases. The matrix of the measurement conducted is indicated in Table 3.1. Three different phases of 0°, 180°, and 90° are used. On-wafer measurements on 70-nm MOSFETs were conducted using this setup with an LSNA. The resulting waveforms obtained for these three experiments are shown in Figures 3.10(a), 3.11(a), and 3.12(a). Note that the incident waveforms $a_1(t)$ and $a_2(t)$ applied are sinusoidal in all cases. However, nonlinear effects are clearly detected in the distorted reflected waveforms $b_1(t)$ and $b_2(t)$ shown in Figures 3.11(a), and 3.12(a). Note that, for case 1 in Figure 3.10, the device does not turn on and thus presents an open. The device trajectory in state-space (v_{GS}, v_{DS}) is shown in Figures 3.10(b), 3.11(b), and 3.12(b) for the three different excitations. Note that the transistor DC drain current is plotted using a contour plot for both negative and positive drain voltages. The inspection of the state-space trajectory reveals that despite a small detectable asymmetry the operation for this on-wafer device remains mostly symmetric, as predicted by the foundry models.

3.3 Device parasitics

In Section 3.1 we compared measured and simulated results for a MOSFET with a fixed gate length, gate width, and number of fingers. In integrated circuits the circuit designer can change the gate length, gate width, and number of fingers. So the device models

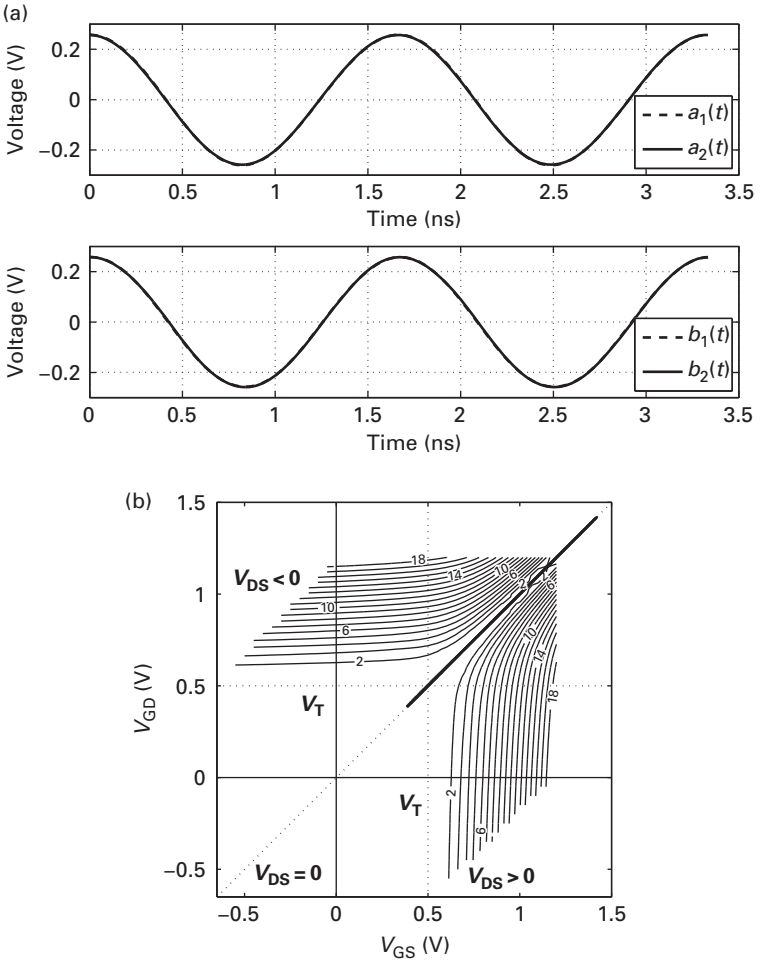


Figure 3.10 (a) Incident and reflected waveforms for case 1 (0°). (b) State-space (v_{GS}, v_{DS}) trajectory with the contour plot of the DC drain current (in mA) superposed.

need to be evaluated as a function of these parameters. In this section we present results for devices with varying gate width.

Given the large quantity of data and waveforms which accumulates as multiple devices of different geometries and different dies are measured, it becomes desirable to introduce some simple figures of merit to evaluate the model fidelity and the process stability. For simplicity the transconductance and threshold voltage are used for this comparison. Note that these parameters are bias-dependent, and we elected here to extract them from the 4-GHz large-signal dynamic transfer characteristic of the device in class-A operation in the saturation region. These parameters are extracted from a least-squares fit of the transfer characteristic as shown in Figure 3.13.

The variation of the transconductance g_m with the gate width W_g for three different dies (dots) and three different models (lines) is shown in Figure 3.14. The measured

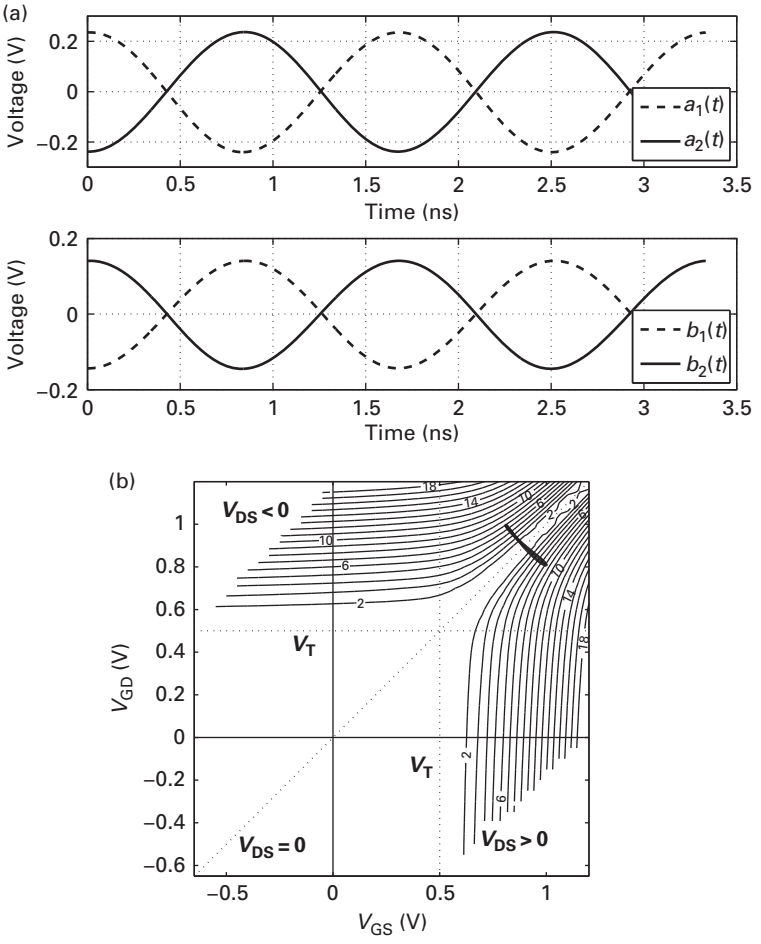


Figure 3.11 (a) Incident and reflected waveforms for case 2 (180°). (b) State-space (v_{GS}, v_{DS}) trajectory with the contour plot of the DC drain current (in mA) superposed.

devices are seen to exhibit a transconductance performance closer to the normal and strong models. A large spread in the measured transconductance is, however, observed at large gate widths.

The variation of the threshold voltage V_{th} with the gate width W_g for three different dies (dots) and three different models (lines) is shown in Figure 3.15. The measured devices are seen to exhibit a threshold voltage between those of the weak and normal models.

The results reported in Figures 3.14 and 3.15 are for the variation of the device performance with the gate width. Similar plots can be generated for the variation of the transconductance and threshold voltage versus gate length L_g and number of gate fingers N_f. The recent trend in compact device modeling for integrated-circuit design is to rely on phenomenological equations with empirical parameters to fit the gate width, gate length, and number of fingers [9]. Alternatively, at high frequencies, the

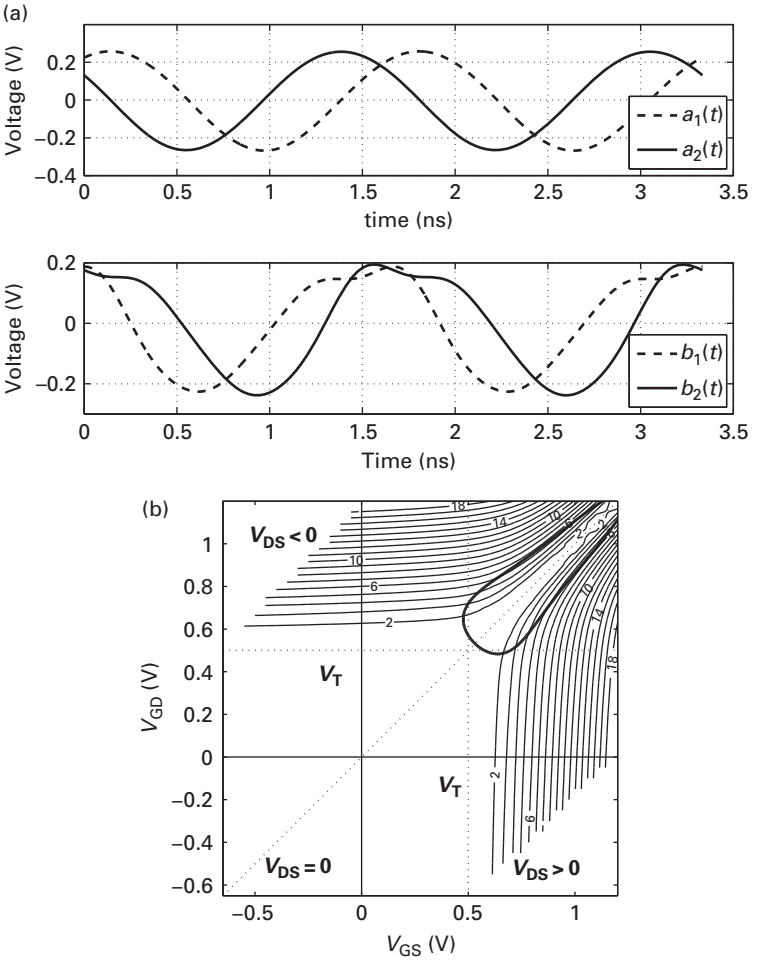


Figure 3.12 (a) Incident and reflected waveforms for case 3 (90°). (b) State-space (v_{GS}, v_{DS}) trajectory with the contour plot of the DC drain current (in mA) superposed. (From [6] with permission, ©2009 IEEE).

physical modeling of the device layout might be desirable in order to optimize the device performance.

The equivalent circuit shown in Figure 3.16 can be used to model the device dependence on the gate width W_g for relatively small gate width [10] [11]. Note the presence of the negative resistance $-R_X/6$ and negative inductance $-L_X/6$, with X standing for any of the terminals S, D, G or B. These negative resistances and negative inductances are needed in order to establish a series feedback when the gate-to-source excitation and the drain-to-source excitations are not applied on the same side of the device; that is, for example, they are applied on side 1 and side 2, respectively. Note, however, that the total input resistance and inductance contributed by each distributed parasitic network remain positive:

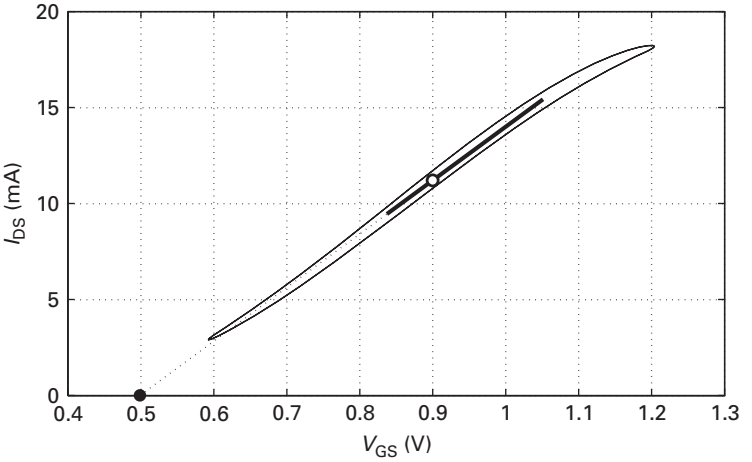


Figure 3.13 Methodology used for extracting an effective transconductance g_m and threshold voltage V_{th} for a large-signal RF excitation in class A.

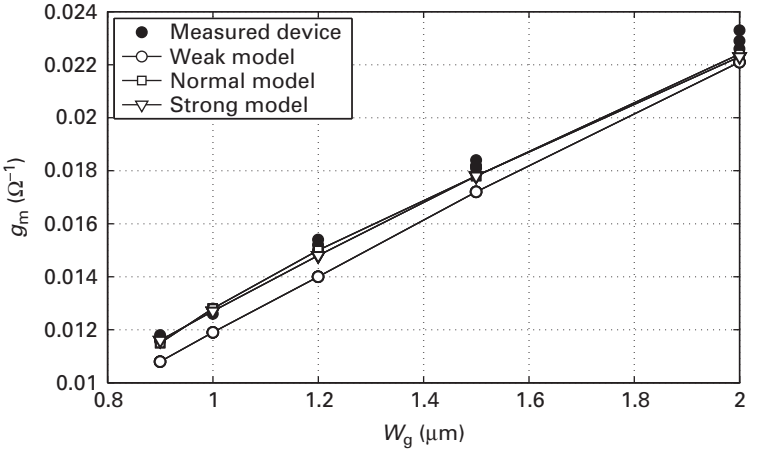


Figure 3.14 Variation of the transconductance g_m with the gate width W_g for three different dies and three different models.

$$\frac{R_X}{2} - \frac{R_X}{6} = \frac{R_X}{3}, \qquad \frac{L_X}{2} - \frac{L_X}{6} = \frac{L_X}{3}.$$

The resistance $R_X/3$ and inductance $L_X/3$ correspond to the well-known result when the excitations V_{GS} and V_{DS} are on the same side of the device [12]. An exact small-signal solution is also available for arbitrary gate widths W_g in Ref. [10] for the case of the three-terminal device. However, W_g is usually kept small enough to avoid distributed effects that degrade the power gain. An exception would be if the parasitics were to be synthesized so as to implement a traveling-wave amplifier. However, in the normal case, the equivalent circuit shown in Figure 3.16 (or multiple sections of it), can usually be

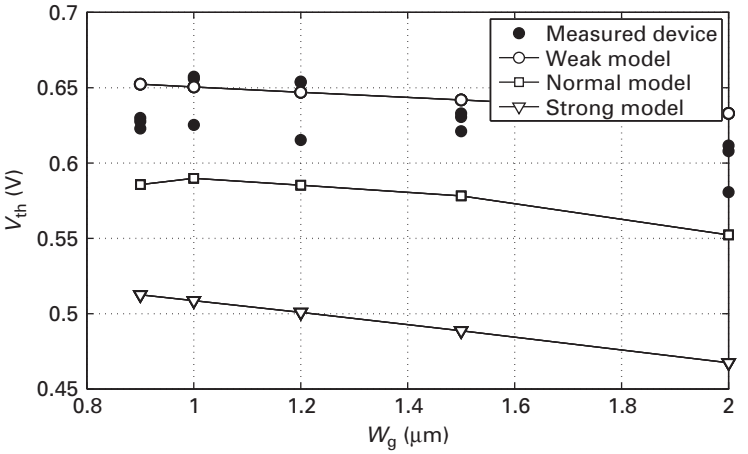


Figure 3.15 Variation of the threshold voltage V_{th} with the gate width W_g for three different dies and three different models.

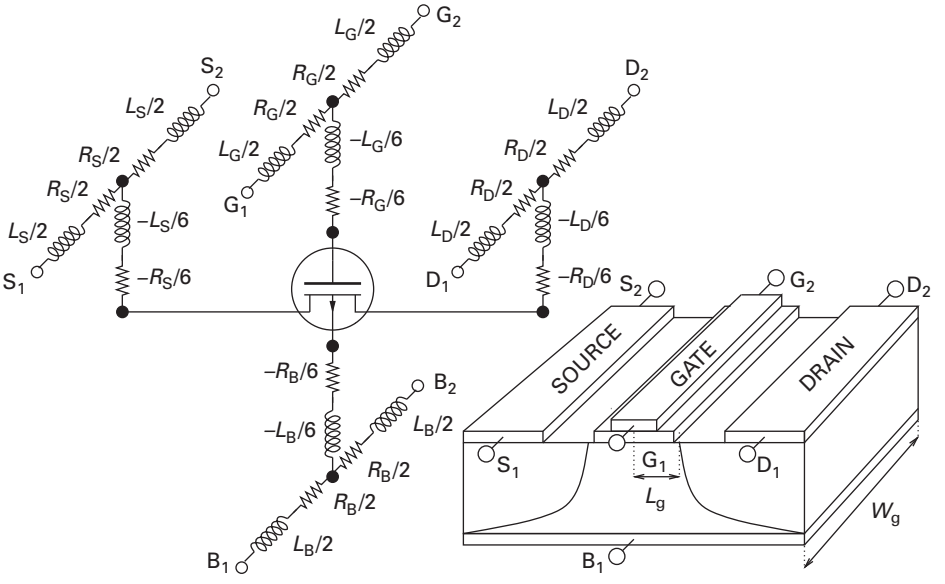


Figure 3.16 Equivalent circuit for an eight-terminal MOSFET for short gate width W_g.

used to model a single transistor-finger in complex multifinger layouts including circular transistors.

Note that the common-gate layout can also facilitate the extraction of the substrate parasitics, if the DC gate bias is selected such that the device is off for all v_{SG} and v_{DG} excitations. When the device is expected to be symmetric, any asymmetry in the measured waveforms could potentially flag a layout error. Indeed, care must be taken, particularly at high frequencies, to independently ground the common terminals on *both* sides, 1 and 2, of the device. Otherwise no well-defined ground return path is provided

for the return current, and a ground loop is introduced by the layout [6]. In such a case the device is then effectively excited by the RF sources with one terminal connected on side 1 and the other terminal connected on side 2, e.g. G1–S2. That is, the device is then operated in the common mode rather than the differential mode, and the equivalent circuit shown in Figure 3.16 is no longer applicable.

3.4 Model extraction from power-sweep measurements

As mentioned in Section 3.1, the NVNA permits one to acquire both the phase and the amplitude of the harmonics. Figure 3.17 shows an example of a power sweep at 4 GHz acquired with an LSNA for a 70-nm MOSFET operated in class A. The dynamic range is seen to be of about 70 dB, as indicated by the fourth harmonic, which is nearly constant below −22 dBm input power. Better dynamic ranges are achieved with the more recent mixer-based NVNAs.

The phases and amplitudes for the various measurements are shown in Figure 3.18. The current and voltage waveforms and dynamic loadlines reconstructed from the amplitude and phase of the harmonics are shown in Figure 3.19, together with the device *I*–*V* characteristics.

Following Ref. [13] a physical model of the device can be implemented using a Volterra expansion. In the example given below, the Volterra expansion is centered on the DC operating point. A conventional pi FET model topology with intrinsic nonlinear capacitances $C_{GS}(v_{GS})$, $C_{GD}(v_{GD})$, and $C_{DS}(v_{DS})$ and with a nonlinear current source $i_D(v_{GS}, v_{DS})$ is used. A third-order expansion is used for the intrinsic voltage dependence of the intrinsic drain current and intrinsic charges. The pad parasitics used for the on-wafer measurement are separately characterized (see the next section for a schematic outline) and removed from the RF measurements. The series source, drain,

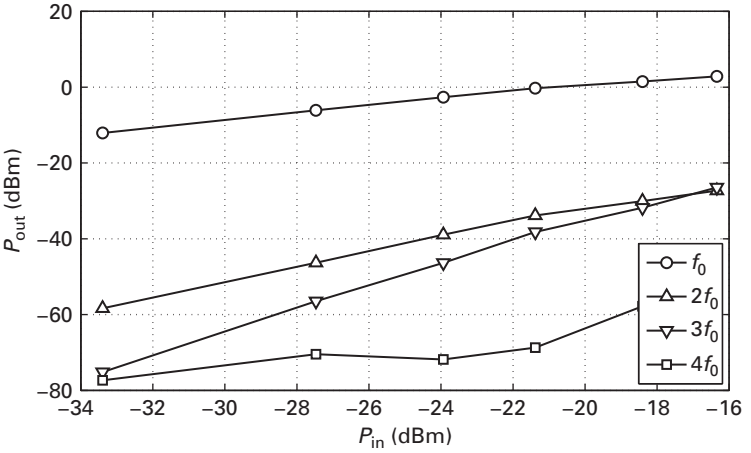


Figure 3.17 Power sweep for a MOSFET operated in class A.

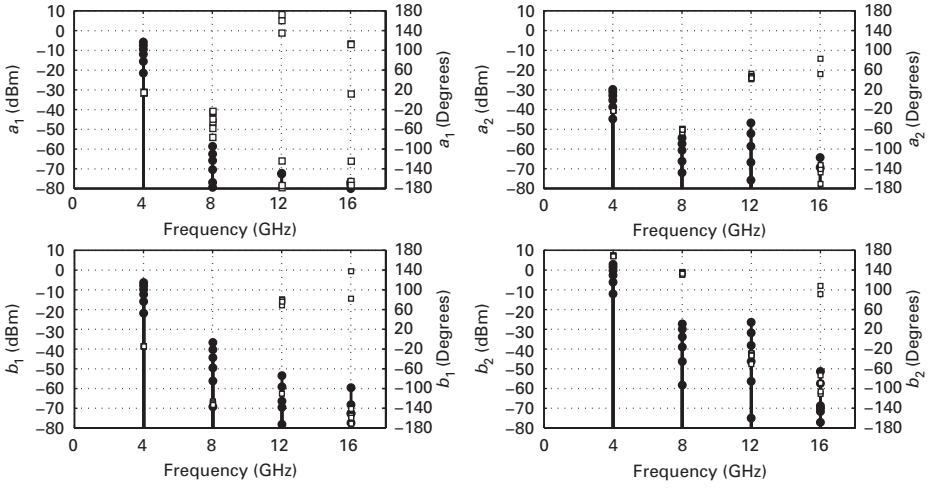


Figure 3.18 Phase and amplitude of the first four harmonics of a_1, b_1, a_2, and b_2 for different input power levels.

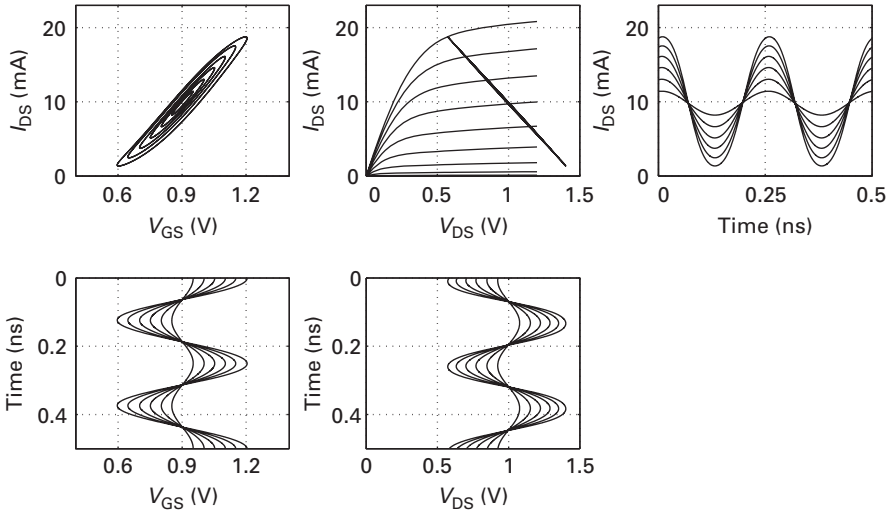


Figure 3.19 Current and voltage waveforms, dynamic transfer characteristics, and dynamic loadlines reconstructed from the measured harmonics in Figure 3.18.

and gate resistances and inductances are also extracted in the analysis. The results for the Volterra physical model obtained from the simultaneous fit of different input power levels are shown in Figure 3.20 for the input dynamic loadlines $i_G(t)$ versus $v_{GS}(t)$ and in Figure 3.21 for the dynamic transfer characteristics $i_D(t)$ versus $v_{GS}(t)$. Clearly a relatively accurate representation of the data is possible with a Volterra series expansion and a simple circuit topology. Note that the simultaneous extraction of such a model for several frequencies is needed in order to extract reliable parasitics.

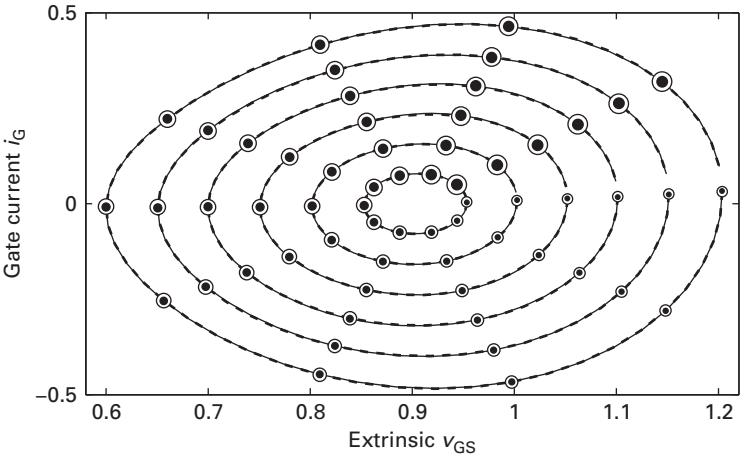


Figure 3.20 Gate current versus gate-to-source voltage at 4 GHz for various input power levels. Measured data (plain lines and black dots) and modeling results (dashed lines and white circles) are compared. Circles of increasing size indicate the time direction in the RF cycle.

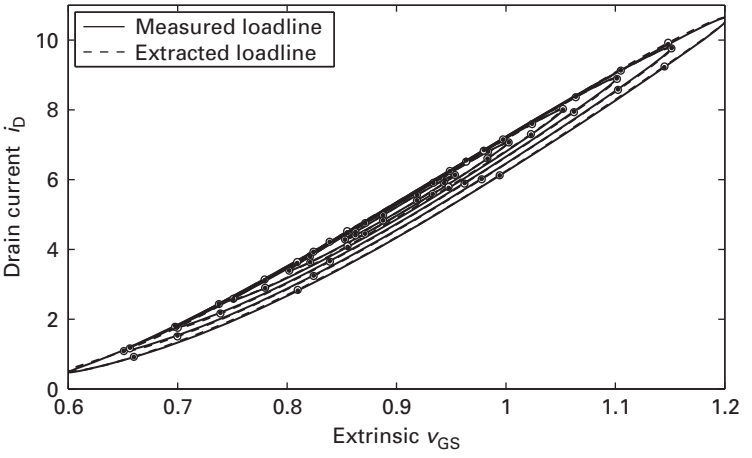


Figure 3.21 Drain current versus gate-to-source voltage at 4 GHz for various input power levels. Measured data (plain lines and black dots) and modeling results (dashed lines and white circles) are compared.

3.5 Model extraction from dynamic loadline measurements

The modeling described in the previous section is limited to a single operating point and for a constant drain load. To map a wider range of operating voltages, one could vary the DC drain voltage as shown in Figure 3.22. This does not, however, account for the potential contribution of memory effects (thermal effects, traps) to be discussed in Chapter 5. Indeed, the device temperature and trap occupation may assume different steady-state values for the various DC drain voltages. When the device is to be operated at a well-defined operating point, the load impedance at the drain terminal can then be

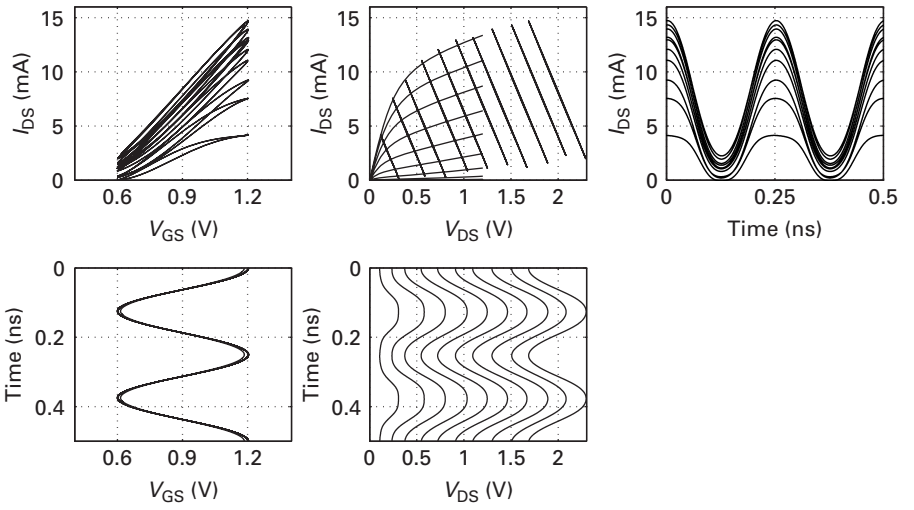


Figure 3.22 Acquisition of large-signal data for multiple drain voltage biases.

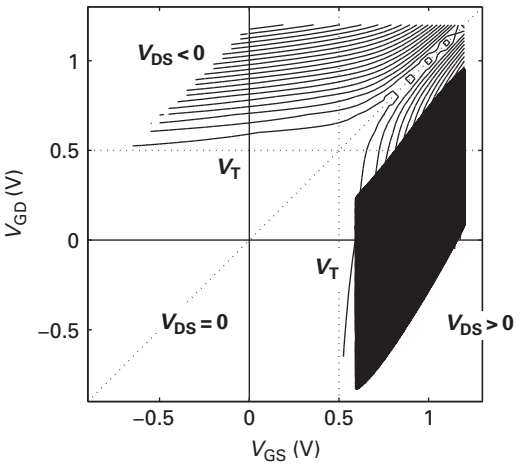


Figure 3.23 Portion of the state-space sweep at 600 MHz acquired in a single RTALP measurement. The modulation frequency is 200 kHz. Only 80 of the 3000 RF cycles taking place in the baseband period are actually shown.

changed using a load tuner. An alternative approach was reported in Ref. [14], in which a multisine (see Chapter 1) is used to map a wide range of drain and gate voltages. This approach is well suited for devices with short memory effects. Yet another approach relies on the use of the dynamic loadline obtained in the real-time active load-pull (RTALP) measurement [15] [16]. The setup for the RTALP measurement will be presented in Chapter 5. An interesting by-product of this technique is that, in a single LSNA measurement record, the RTALP measurement generates a dynamic loadline that sweeps a wide range of gate and drain voltages, as is demonstrated in Figures 3.23 and 3.24.

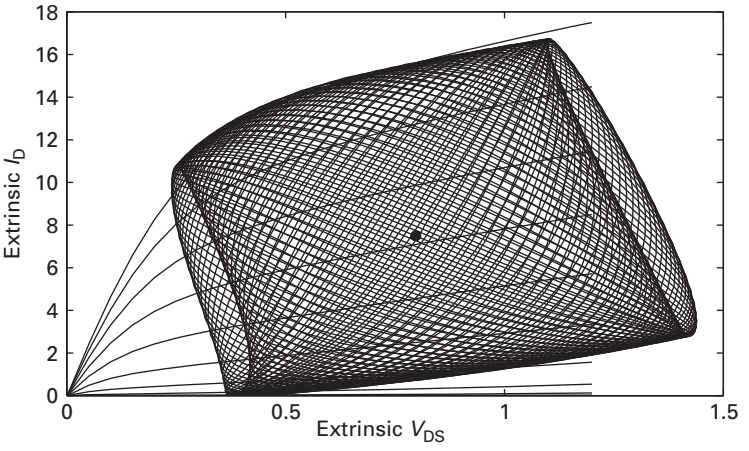


Figure 3.24 Portion of the dynamic loadline at 600 MHz acquired in a single RTALP measurement. Only 80 of the 3000 RF cycles taking place in the baseband period are shown. Superposed on the graph is the extrinsic *I*–*V* characteristics (plain lines) and the DC bias point (black dot).

Figure 3.23 shows the locus of the gate-to-source and gate-to-drain voltage swept in a single RTALP measurement (10 ms data acquisition). The frequency of the input signal is 600 MHz and the frequency of the output signal is 600.2 MHz. The RF dynamic loadline is periodic and repeats at the modulation frequency of 200 kHz, while slowly sweeping the voltage space in a single baseband period. Note that only 80 of the 3000 RF cycles taking place in the dynamic loadline period are actually plotted. These 80 RF cycles are selected to be equally spaced on the baseband period. Figure 3.24 shows the resulting dynamic loadline superposed on the extrinsic *I*–*V* characteristics. The gate voltage of the *I*–*V* characteristics varies from 0 to 1.2 V in steps of 0.1 V. A single LSNA measurement is seen to be capable of capturing a wide range of the *I*–*V* characteristics. The availability of such a wealth of data from a single large-signal RTALP measurement suggests that these data could be used for directly extracting a large-signal device model. Indeed, the direct extraction of a device model from such measurements has been reported [17] [18] and offers some exciting possibilities for further development. This approach is further explored below for the RTALP reported in Figure 3.24.

Figure 3.25 shows the extracted intrinsic loadline (plain line) and the DC bias point (large black dot) after the parasitics have been removed. The equivalent circuit used for the parasitics is shown in Figure 3.26. It includes the on-wafer pads and the gate, source, and drain series resistances. Note that in Figure 3.25 only 20 of the 3000 RF cycles in the dynamic loadline period are now presented to facilitate the comparison between the modeled (plain line) and measured dynamic loadlines (solid line). The intrinsic device model relies on the charge topology shown in Figure 3.9, except that a quasi-static approximation ($\tau_G = \tau_D = 0$) is used for this low frequency of operation. A B-spline representation (see [19]) is used to extract the voltage dependence of the current and

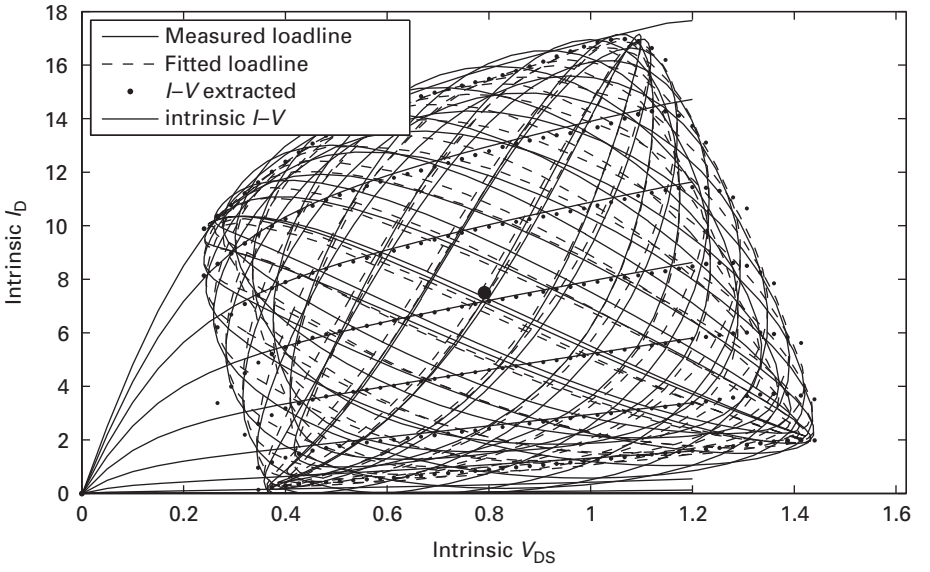


Figure 3.25 Comparison of the measured (plain line) and modeled (dashed line) dynamic loadline. Superposed are the extracted (dotted line) and DC intrinsic (plain line) *I–V* characteristics. Only 20 of the 3000 RF cycles taking place in the baseband period are shown. The DC bias point is indicated with a large black dot.

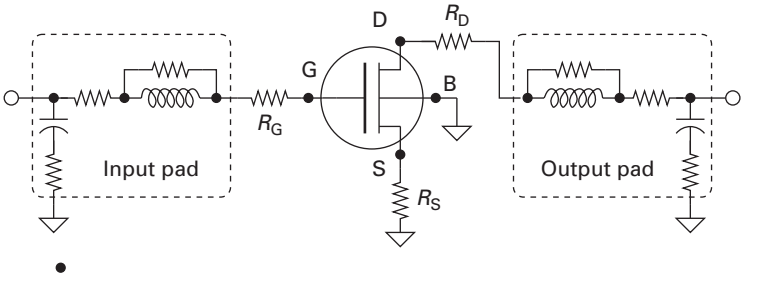


Figure 3.26 Model used to recover the intrinsic FET dynamic loadline and DC *I–V* characteristics.

charges. The intrinsic *I–V* characteristic, gate charge, and drain charge are simultaneously extracted by fitting the input and output dynamic loadlines. An approximate fit of the output loadline is obtained with this model. For a wide range of data points, the extracted intrinsic *I–V* characteristics (small dots) are consistant with the *I–V* characteristic (plain lines) measured at DC, once the resistive parasitics have been accounted for. However, the extracted intrinsic *I–V* characteristics depart noticeably from the measured ones on the edge of the loadline trajectory.

References

[1] J. Verspecht and D. Schreurs, "Measuring transistor dynamic loadlines and breakdown currents under large-signal high-frequency operating conditions," in *1998 IEEE MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1495–1498, 1998.

[2] D. Schreurs, L. Pantisano, and B. Kaczer, "Analysing impact of MOSFET oxide breakdown by small- and large-signal HF measurements," in *64th ARFTG Microwave Measurements Conference*, pp. 85–91, 2004.

[3] D. Schreurs, E. P. Vandamme, and S. Vandenberghe, "Capabilities of vectorial large-signal measurements to validate RF large-signal device models," in *58th ARFTG Conference Digest*, pp. 1–6, 2001.

[4] X. Xi, M. Dunga, J. He, W. Liu, K. M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu, *BSIM4.3.0 MOSFET Model*, The Regents of the University of California, 2003.

[5] V. Balasubramanian, "Characterization of Sub-90 nm Gate Length RF MOSFETs Using Large Signal Network Analyzer," Thesis, Master of Philosophy, Ohio State University, 2008.

[6] P. Roblin, "Application of large signal network analyzer measurements to model verification and device modeling," in *10th Annual Wireless and Microwave Technology Conference, WAMICON '09*, pp. 1–5, 2009.

[7] Jianjun Xu, D. Gunyan, M. Iwamoto, J. M. Horn, A. Cognata, and D. E. Root, "Drain–source symmetric artificial neural network-based FET model with robust extrapolation beyond training data," in *2007 IEEE/MTT-S International Microwave Symposium*, pp. 2011–2014, 2007.

[8] K. Yhland, N. Rorsman, M. Garcia, and M. Merkel, "A symmetrical nonlinear HFET/MESFET model suitable for intermodulation analysis of amplifiers and resistive mixers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 48, No. 1, pp. 15–22, Jan. 2000.

[9] B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill Higher Education, 2001.

[10] S. Lee, P. Roblin, and O. Lopez, "Modeling of distributed parasitics in power FETS," *IEEE Transactions on Electron Devices*, Vol. 49, No. 10, pp. 1799–1806, Oct. 2002.

[11] S. Lee and P. Roblin, "Modeling of distributed parasitic effects in FETs and new 4-port equivalent circuit for small gate-width FET building blocks," in *Proceedings of International VLSI Multilevel Interconnection Conference*, Marina Del Rey, CA, 2003.

[12] P. Wolf, "Microwave properties of Schottky-barrier field-effect transistors," *IBM Journal of Research and Development*, Vol. 14, pp. 138–139, Mar. 1970.

[13] J. C. Pedro and J. Perez, "Accurate simulation of GaAs MESFET's intermodulation distortion using a new drain–source current model," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 42, No. 1, pp. 25–33, Jan. 1994.

[14] D. Schreurs and K. A. Remley, "Use of multisine signals for efficient behavioral modeling of RF circuits with short-memory effects," in *61st ARFTG Conference Digest*, pp. 65–72, 2003.

[15] F. Verbeyst and M. Vanden Bossche, "Real-time and optimal PA characterization speeds up PA design," in *34th European Microwave Conference Digest*, Amsterdam, pp. 431–434, 2004.

[16] X. Cui, S. J. Doo, P. Roblin, G. H. Jessen, R. G. Rojas, and J. Strahler, "Real-time active load-pull of the 2nd & 3rd harmonics for interactive design of non-linear power amplifiers, in *68th ARFTG Conference Digest*, Colorado, 2006.

[17] M. C. Curra-Francos, P. J. Tasker, M. Fernandez-Barciela, Y. Campos-Roca, and E. Sanchez, "Direct extraction of nonlinear FET *Q–V* functions from time domain large signal measurements," *IEEE Microwave and Guided Wave Letters*, Vol. 10, No. 12, pp. 531–533, Dec. 2000.

[18] D. Schreurs, J. Verspecht, S. Vandenberghe, and E. Vandamme, "Straightforward and accurate nonlinear device model parameter-estimation method based on vectorial large signal measurements," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 50, No. 10, pp. 2315–2319, Oct. 2002.

[19] P. Roblin and H. Rohdin, *High-Speed Heterostructure Devices*, Cambridge University Press, 2002.

4 Characterization and modeling of memory effects in RF power transistors[1]

This chapter discusses memory effects exhibited by transistors and how they affect their large-signal RF performance. Memory effects include self-heating, traps, and parasitic bipolar effects in SOI-MOSFETs. Distributed and transient thermal models will be discussed first. Large-signal measurement techniques for characterizing memory effects in transistors using pulsed biased and pulsed-RF large-signal measurements will be presented. Results from combined deep-level optical spectroscopy and large-signal measurement will then be introduced. Finally, the correlation between trapping and noise will be discussed. The chapter will conclude then with a discussion of the cyclostationary effect, according to which the average trapping population and device noise characteristics can be altered by the fast RF signals under large-signal RF operation.

4.1 Importance of memory effects in RF devices

GaN HEMTs provide a good example of devices strongly affected by memory effects. High-electron-mobility transistors (HEMTs) are among the most successful heterostructure three-terminal devices to have emerged over the last couple of decades [1]. GaN-based HEMTs show high transconductance, high cutoff frequencies, and good thermal management such that they are suitable for high-power and high-speed applications with minimal cooling [2] [3].

One of the obstacles for GaN HEMTs is current collapse or knee walk-out, which is known to result from the effects of surface trap states [4] [5] [6]. Many studies concerning the reduction of current collapse using various device processing techniques such as surface passivation, modified buffer layer designs, and field plates have been reported [3] [7]. The various memory effects at play depend on the substrate and the fabrication techniques used. Various thermal and trap measurement techniques will be discussed in this chapter to characterize these memory effects with a specific emphasis on large-signal RF characterization techniques.

[1] Research collaboration with Wen Hua Dai, Seok Joo Doo, Chieh Kai Yang, Inwon Suh, Aaron Arehart, Andrew Malonis, and Stephen Ringel is gratefully acknowledged.

4.2 Distributed and transient models for self-heating in power transistors

4.2.1 Steady-state thermal modeling

Transistors, like any other device operated under non-equilibrium conditions, will dissipate power. Non-equilibrium occurs when energy is applied, be it in the form of a DC bias, RF voltages, a temperature gradient or electromagnetic waves. Owing to the finite thermal conductivity of materials, the power dissipated, P_{diss}, in the device will in turn increase the device temperature by ΔT_{dev}. This process is referred to as self-heating and is measured by determining the thermal resistance:

$$R_{\text{th}} = \frac{\Delta T_{\text{dev}}}{P_{\text{diss}}}.$$

As a result of self-heating, the device temperature can be quite different from the substrate temperature T_{sub}:

$$T_{\text{dev}} = T_{\text{sub}} + R_{\text{th}} \times P_{\text{diss}},$$

and the device characteristics change. For example consider the *I–V* characteristics of an LDMOSFET shown in Figure 4.1. The measured [1] average surface temperature of the LDMOSFET is superposed on the *I–V* characteristics. The substrate temperature is

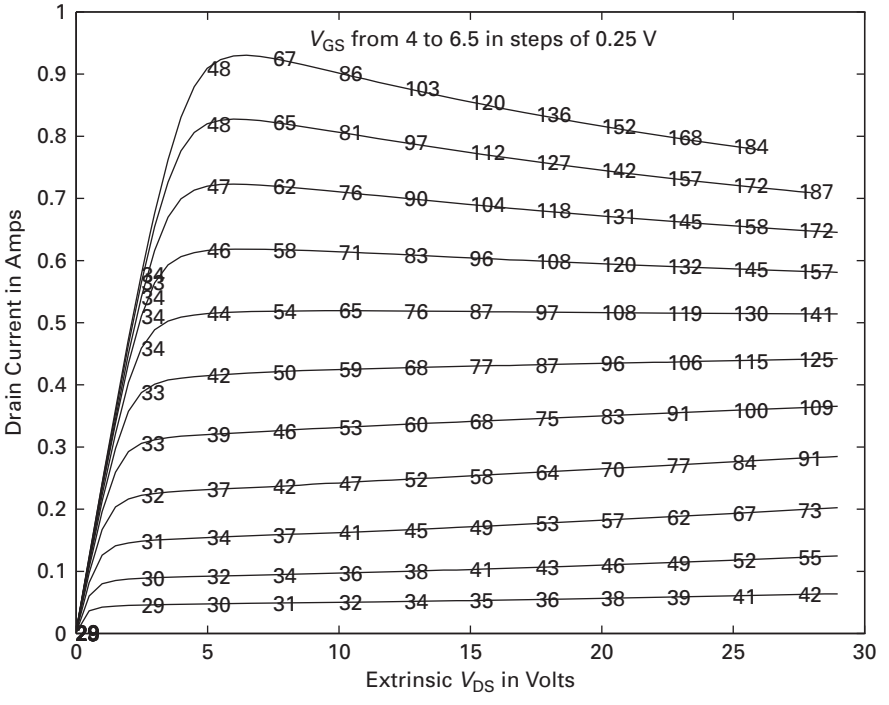


Figure 4.1 Measured *I–V–T* (electrothermal characteristics) of an LDMOSFET for a constant substrate temperature of 29 °C. (Measured by Siraj Akhtar at Ohio State University.)

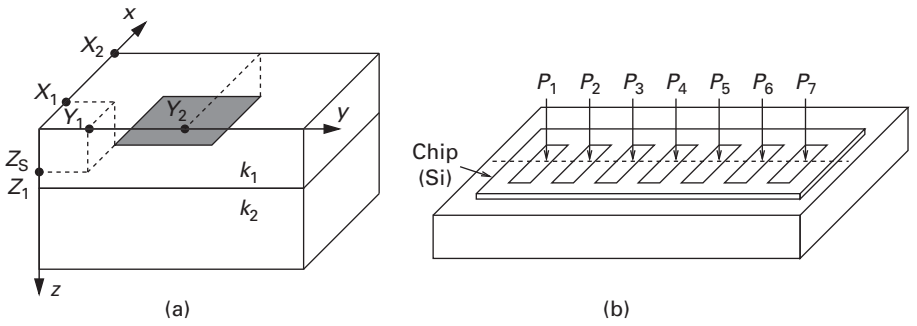


Figure 4.2 Two-layer heat transfer (a) and parallel heat sources (b).

maintained at 29 °C. Clearly the surface temperature can depart substantially from the substrate temperature. At high gate voltages the increase in device temperature is seen to decrease the drain current due to the reduction of the electron mobility and saturation velocity at high temperatures. At lower gate voltages, the drain current usually increases in the LDMOSFET due to the reduction of the threshold voltage.

Obviously, reducing the three-dimensional temperature distribution inside the device to a single average temperature is an approximation. However, for field-effect transistors, which are surface devices, monitoring the surface is usually sufficient. Figure 4.2(b) shows a simplified multi-cell LDMOSFET structure and Figure 4.2(a) shows a single cell. As shown in Figure 4.2(a) the device structure considered consists of a silicon layer, with the active areas (modeled as rectangular surface heat source) sitting on top, and of a copper layer corresponding to the package. Many methods to calculate the steady-state temperature distribution in such layered structures have been reported. Numerical methods such as FEM [8] and FDM [9] are powerful techniques and are often used to analyze complex structures. The image method [10] [11] provides an alternative and accurate approach for simpler multiple-layer structures. It relies on the summation of a truncated series of Green functions that are solutions of three-dimensional (3D) heat-transfer problems of image heat sources in infinite and uniform media. The steady-state Green function (Equation (4.2)) is obtained from the Poisson equation (Equation (4.1))

$$\nabla^2 T_{\mathrm{G}} = -\frac{1}{k}\delta(\mathbf{r} - \mathbf{r}'), \tag{4.1}$$

$$T_{\mathrm{G}}(\mathbf{r}, \mathbf{r}') = \frac{1}{4\pi k|\mathbf{r} - \mathbf{r}'|}, \tag{4.2}$$

where k is the thermal conductivity, and $\mathbf{r}$ and $\mathbf{r}'$ are the locations of the observation point and the heat source, respectively.

An analytic solution for the temperature increase for a surface heat source of two-dimensional (2D) heat density q_{s} located at $z = Z_{\mathrm{s}}$ can then be obtained by integrating the Green function from (X_1, Y_1) to (X_2, Y_2) (see Ref. [11]). The half-space solution for the two-layer system shown in Figure 4.2 is then obtained using the image method described in Ref. [11]. The image method enforces the adiabatic boundary condition (no

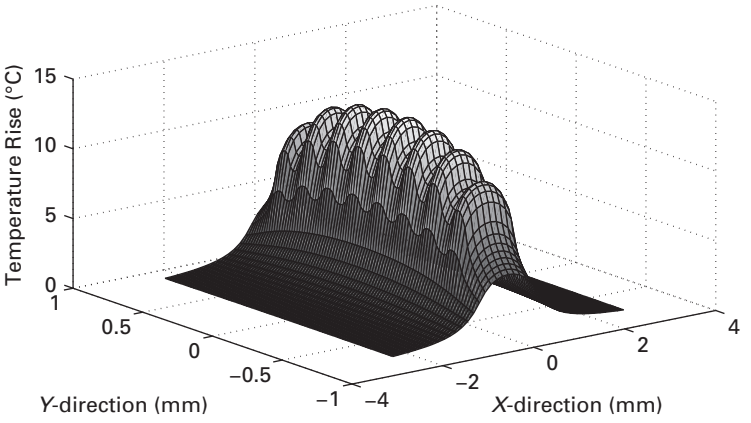


Figure 4.3 Two-dimensional distribution of the surface temperature at the surface of an eight-cell LDMOSFET, $P = 1$ W per element. (From [12] with permission, ©2003 IEEE.)

convection) at the surface and the boundary conditions at the interface of layers 1 and 2, which are given by

$$T(z = z_1^-) = T\left(z = z_1^+\right), \tag{4.3}$$

$$-k_1 \left.\frac{\partial T}{\partial z}\right|_{z=z_1^-} = -k_2 \left.\frac{\partial T}{\partial z}\right|_{z=z_1^+}. \tag{4.4}$$

Using this model the surface temperature distribution was calculated for an eight-cell transistor (see Figure 4.3). One can see that the temperature peaks at the center of each cell and that the fingers on the edge are cooler due to the lateral heat flow. A qualitative comparison of the predicted 2D surface temperature with the measured infrared thermal imaging is given in [12].

4.2.2 Implementation of the distributed thermal model

To account for distributed effects in an LDMOS transistor the temperature of each transistor cell can be computed and fed back to the electrical FET model. This strategy is shown in Figure 4.4, where the primary device consists of N parallel cells. All cells have the same electrical model, and their temperatures vary depending on their locations within the chip. The mutual thermal resistances R_{ij} account for the heat transferring horizontally between cells. The image method can be used to find the self and mutual R_{th} of Equation (4.5):

$$\begin{bmatrix} \Delta T_1 \\ \Delta T_2 \\ \vdots \\ \Delta T_n \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ R_{21} & R_{22} & \dots & R_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ R_{n1} & R_{n2} & \dots & R_{nn} \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_n \end{bmatrix}. \tag{4.5}$$

Table 4.1. Averaged R_{th} calculated using the image method

Gate width (mm)	Number of subsets	R_{th} (°C/W)
3	1	6.17
12	4	2.79
48	16	1.07
83.4	26	0.74

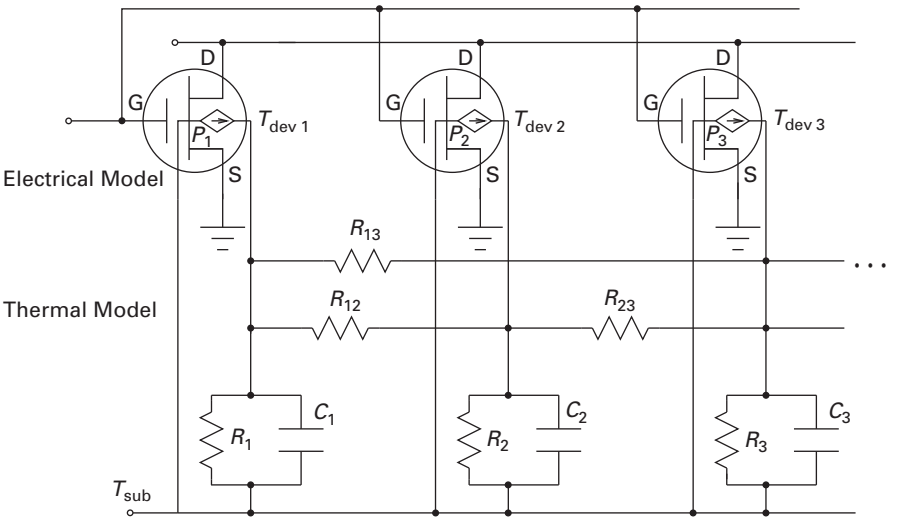


Figure 4.4 The electrothermal equivalent circuit for a multi-cell LDMOSFET.

Although the distributed model is attractive for modeling each cell of the transistor and predicting possible runaway situations, in most RF simulations reducing the device to a single average temperature and a single thermal resistance is usually sufficient for accurately predicting the collective response of multifinger devices to CW and modulated RF excitations [12].

The average thermal resistance can be expected to reduce with increasing device area. A simplified scaling rule might assume that the average thermal resistance R_{th} is inversely proportional to the device active area. However, this is not accurate because the lateral heat flow is not accounted for. This is demonstrated in Table 4.1 which gives the average R_{th} of devices of various areas calculated with the image model discussed in the next section. Another consequence of the vertical and horizontal heat flow is the presence of a temperature gradient between the center and edge cells. The electrical characteristics are in turn affected by the temperature distribution and therefore the cells in the transistor behave non-uniformly.

The average R_{th} value for a 12-mm device [13] was measured to be 2.80 °C/W in an amplifier testbed, which is consistent with the image-method calculation reported in Table 4.1. Note that the thermal resistance R_{th} obtained in the experimental setup [14]

is an average value that depends on the size of the measured area under the infrared thermometer (see Figure 6 in Ref. [12]).

The distributed model can be extended by adding the thermal capacitance of each cell. However, the thermal transient response of the distributed model closely follows that of a single *R–C* time-constant response whereas, as we shall see in the next section, the device exhibits a multiple-time-constant thermal-transient response. The distributed model shown in Figure 4.4 is therefore not the best suited to model the broadband thermal response of the device and an alternative approach is pursued in the next section.

4.2.3 Transient thermal response

We shall now investigate the transient thermal response of a transistor when a step power dissipation is generated. We start again with the problem of an infinite and uniform medium. Its associated Green function is the solution of the time-dependent heat-flow equation:

$$\frac{\partial T_{\mathrm{G}}}{\partial t} - D\,\nabla^2 T_{\mathrm{G}} = \frac{1}{\rho C_p}\delta(\bar{\mathbf{r}} - \bar{\mathbf{r}}')\delta(t - t'), \tag{4.6}$$

where $D = k/(\rho C_p)$ is the thermal diffusivity, C_p the thermal capacity, and ρ the density. The Green-function solution of Equation (4.6) is a three-dimensional Gaussian, which is given in Ref. [15]. Integrating the Green function over time and over the rectangular surface heat source yields the temperature step-response solution (see Ref. [16])

$$T_{\mathrm{step}}(x, y, z, t) = \frac{q_{\mathrm{s}}}{4k\sqrt{\pi}} \int_0^{2L(t)} \exp\left(-\frac{z^2}{u^2}\right) \left[\mathrm{erf}\left(\frac{X_2 - x}{u}\right) - \mathrm{erf}\left(\frac{X_1 - x}{u}\right)\right] \times \left[\mathrm{erf}\left(\frac{Y_2 - y}{u}\right) - \mathrm{erf}\left(\frac{Y_1 - y}{u}\right)\right] du, \tag{4.7}$$

where X_1, X_2, Y_1, and Y_2 are the coordinates of the corners of the rectangular source, $L(t) = \sqrt{D\,t}$ is the thermal diffusion length, and $\mathrm{erf}(x)$ is the error function.

One can account for multiple fingers by using the superposition principle. Figure 4.5 plots the change in temperature distribution with time in an eight-cell transistor along the dashed line in Figure 4.2(b).

Equation (4.7) is the step-response solution for a rectangular heat source in a half-space uniform medium when the heat source is at the surface ($Z_{\mathrm{s}} = 0$ is not a necessary assumption). We now need to extend this transient solution to layered structures. Unlike for the steady-state case, an exact transient solution cannot be obtained by the image method for the general case. Indeed, because of the difference in heat diffusion between each layer and the next, the boundary conditions involved (see Equations (4.3) and (4.4)) cannot be satisfied by the image solution at all times and all positions across the various interface planes. The image method can, however, be approximately applied to the calculation of the transient response in the two-layer system [17].

Two approximate approaches are first considered in order to obtain a solution satisfying the boundary conditions of Equations (4.3) and (4.4). In the first approach all

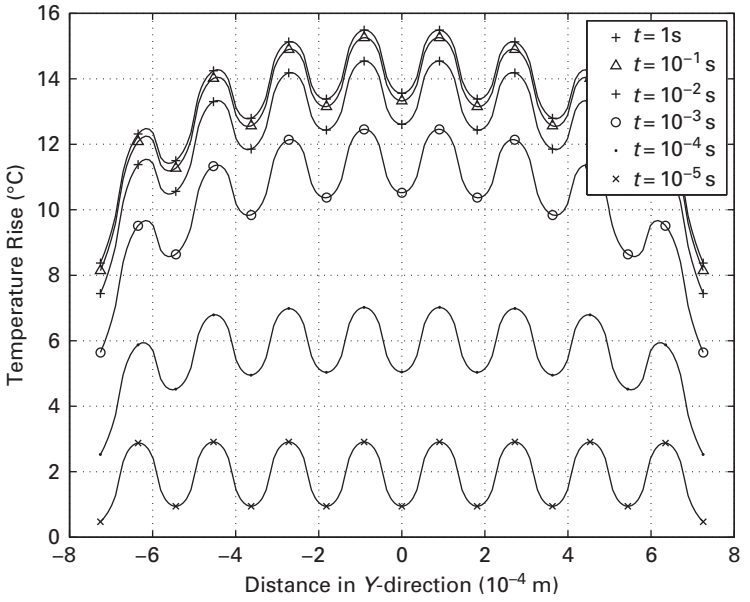


Figure 4.5 Temperature distribution vs. time in silicon. Each of the eight heat sources is of dimensions $1270\,\mu\text{m} \times 90.68\,\mu\text{m}$, with a center-to-center spacing of $2 \times 90.68\,\mu\text{m}$.

layers are assumed to have the same diffusivity D (but with different values of the thermal conductivity k). In the second approach, which is applicable to a two-layer system, the thermal conductivity k in the second layer is assumed to be infinite. The solutions obtained for both approaches satisfy exactly the boundary conditions but yield different results because they are based on different system assumptions.

We now compare the applicability of these system assumptions to the device under study. In the case considered the first layer (silicon) has a thickness of $300\,\mu\text{m}$ with $k_1 = 148$ W/(m K) and $D_1 = 8.47 \times 10^{-5}$ m^2/s, and the second layer (copper) has infinite thickness with $k_2 = 390$ W/(m K) and $D_2 = 1.12 \times 10^{-7}$ m^2/s. The transistor chip sits on an 80-mm-thick Cu–W flange, and the amplifier built in this experiment has a copper heat sinker of thickness 12.5 mm. Its thermal time constant is above 20 minutes. Therefore it is modeled as a half-space medium. These values indicate that the second approach should give the most realistic results.

Figure 4.6 compares the transient temperature rises at the middle of the device calculated with the finite-element method (FEM, exact solution) and the extended image method (approximate solution). The curve of $D = D_1$ gives the fastest temperature response, and the curve of $D = D_2$ gives the slowest response. Note that they both reach the same steady-state temperature, which depends solely on k_1, k_2, and the silicon layer thickness. When assuming $k_2 = \infty$, the temperature follows the $D = D_1$ curve for a short time and then deviates when the heat wave reaches the copper layer. The 3D FEM heat-transfer analysis conducted for this structure yields a similar shape in temperature rise. The FEM curve exhibits a knee at approximately $t = 10^{-3}$ s, corresponding

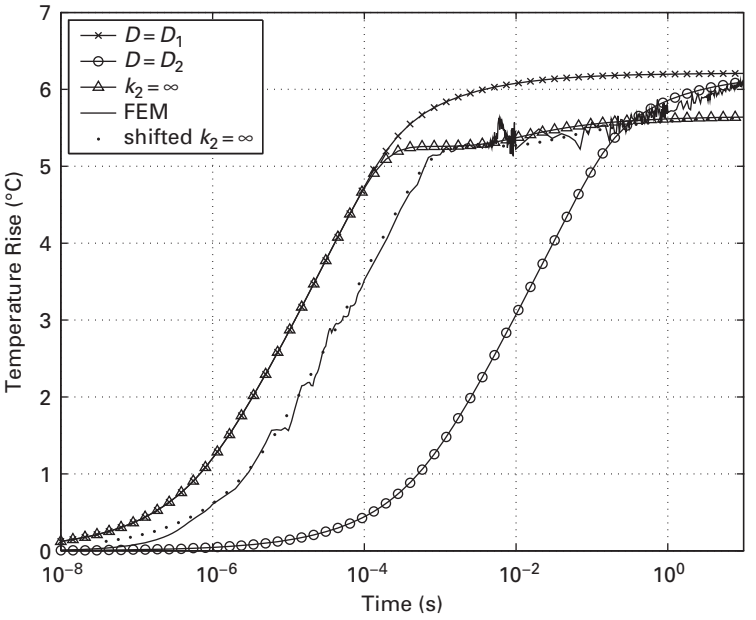


Figure 4.6 Transient temperature rise in a two-layer structure (silicon on copper). The thickness of silicon is 300 μm and the input power is 1 W. (From [12] with permission, ©2003 IEEE.)

to the time required for the heat wave to hit the silicon–copper interface, estimated as $t = L^2/D = 1.06 \times 10^{-3}$ s. The $k_2 = \infty$ (dotted) curve when shifted in logarithmic time scale (time multiplied by $(D_1/D_2)^{1/2}(k_1/k_2)^2$) overlaps with the FEM (plain) curve from time zero up to 1 s. This shows that the shifted $k_2 = \infty$ curve provides a reasonable approximation that can be used in models for circuit simulations.

4.2.4 Modeling of the transient thermal response

The shape of the transient thermal response in Figure 4.6 suggests that it can be modeled by *R–C* circuits like the ones in Figures 4.7 and 4.8, where the instantaneous power is represented by a current source and the temperatures are represented by node voltages. The thermal resistances and capacitances can be extracted by fitting the step response. Figures 4.7 and 4.8 show three fitted curves against the FEM calculation for a 1-W and a 60-W LDMOSFET, respectively. Among the one-stage (RC1), three-stage (RC3), and five-stage (RC5) multi-stage *R–C* circuits, the best fit is achieved with RC5 for both the 1-W and the 60-W device. The values used for the various thermal models for the 60-W LDMOSFET are given in Table 4.2. Similar results have been obtained for SOI-MOSFETs [18] [19], for which the modeling of the thermal response was found to require multiple time-constants. In Chapter 9 we shall make use of these multiple-time-constant thermal models together with a temperature-dependent device model [20] to discuss the impact of thermal memory effects on predistortion linearization.

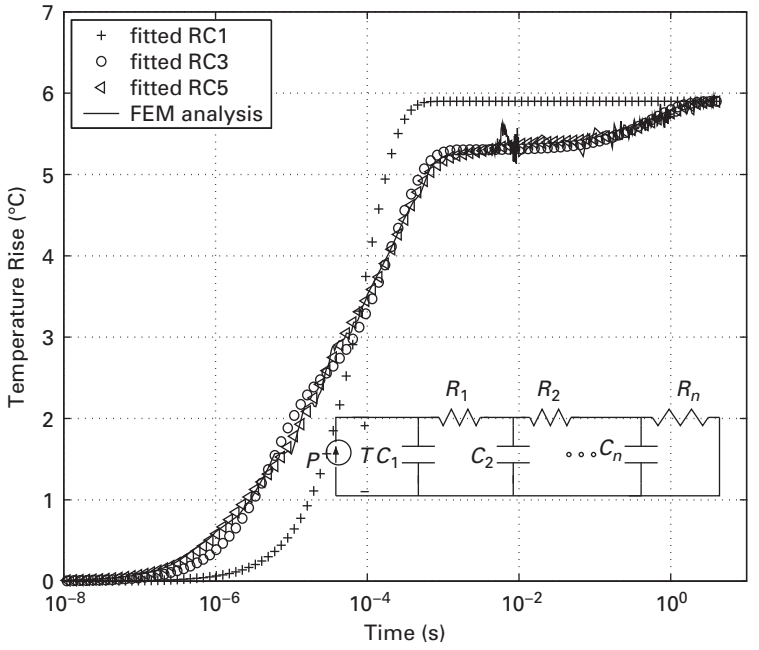


Figure 4.7 Fit of the thermal step response by the RC1, RC3, and RC5 networks for a 1-W LDMOSFET. (From [12] with permission, ©2003 IEEE.)

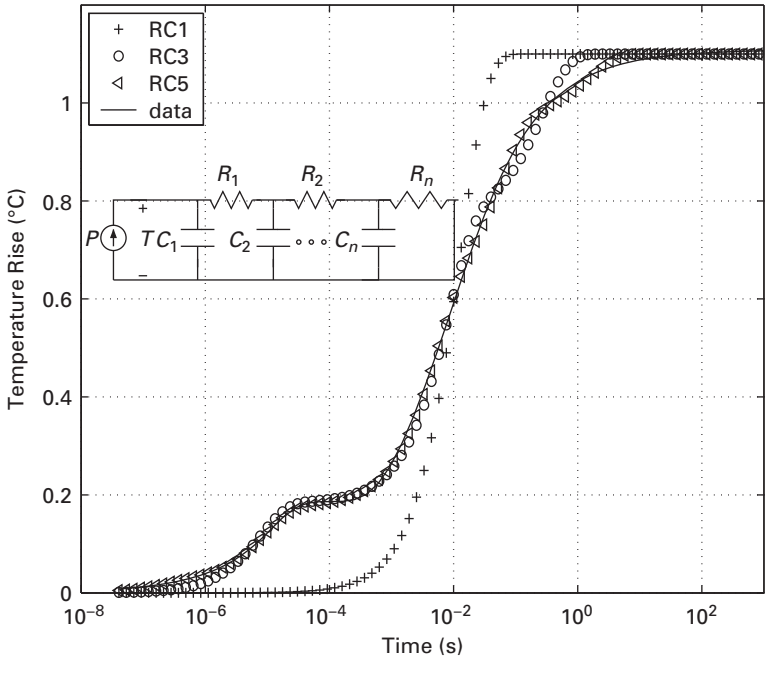


Figure 4.8 Fit of the thermal step response by the RC1, RC3, and RC5 networks for a 60-W LDMOSFET.

Table 4.2. Fitted multi-stage *R–C* thermal parameters for the 60-W LDMOSFET

	R_{th1} (Ω)	R_{th2} (Ω)	R_{th3} (Ω)	R_{th4} (Ω)	R_{th5} (Ω)
RC1	1.10				
RC3	1.846×10^{-1}	6.020×10^{-1}	3.135×10^{-1}		
RC5	3.176×10^{-2}	1.456×10^{-1}	4.791×10^{-1}	3.223×10^{-1}	1.213×10^{-1}
	C_{th1} (F)	C_{th2} (F)	C_{th3} (F)	C_{th4} (F)	C_{th5} (F)
RC1	1.180×10^{-2}				
RC3	4.229×10^{-5}	1.377×10^{-2}	8.523×10^{-1}		
RC5	7.742×10^{-6}	5.811×10^{-5}	1.072×10^{-2}	1.636×10^{-1}	1.237×10^{-1}

The results reported here are for the LDMOSFETs. Note that very fast thermal processes have been reported in AlGaN/GaN HEMTs. Time-constants as small as 15 ns [21] have been used to fit their transient response. Temperature rises of 65 °C and 30 °C in 200 ns have been measured using time-resolved Raman thermography [22]. Clearly, thermal effects could potentially contribute to broadband memory effects. Note that measurements on the variation in time of the large-signal RF gain of AlGaN/GaN HEMTs have detected time variations (memory effects) on the order of microseconds [23]. Multiple recording (see Section 2.9) with an LSNA was used for these measurements.

4.3 Identification of self-heating using pulsed *I–V* pulsed-RF measurements

In this section we shall examine *I–V* knee walk-out in a GaN HEMT fabricated on a sapphire substrate and use NVNA measurements to demonstrate that, for this particular device technology, the *I–V* knee walk-out is of thermal origin. To do so we will use the dynamic loadlines acquired with a large-signal network analyzer (LSNA) to demonstrate how, in pulsed *I–V* pulsed-RF operation, the undesirable *I–V* knee walk-out can be effectively suppressed when keeping the same substrate temperature and then reintroduced when setting the substrate temperature to the surface temperature of the device in CW operation.

The device under consideration is a 0.35 μm AlGaN/GaN HEMT 2 × 150 μm wide, fabricated on a sapphire substrate with Ir/Au Schottky contacts and Ti/Al/Ni/Au ohmic contacts [24]. Figure 4.9 shows the DC *I–V* characteristics of the DUT (plain line) and pulsed *I–V* characteristics (dash–dot line) for $v_{GS} = 0$ V. The large negative drain conductance in the DC *I–V* characteristics (plain line) for $v_{GS} = 0$ V compared with the small positive drain conductance in the pulsed *I–V* characteristics (dash–dot line) for $v_{GS} = 0$ V is typically indicative of a substantial build up of temperature in the DUT.

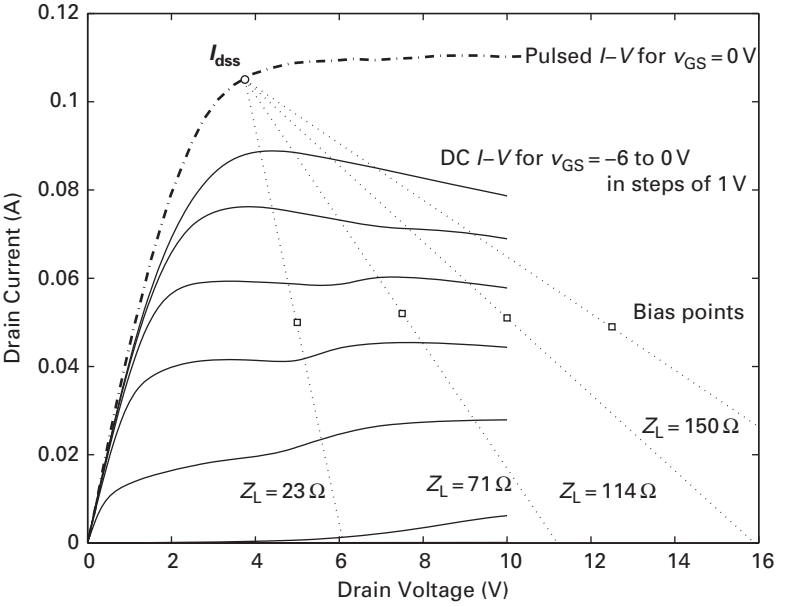


Figure 4.9 DC *I–V* characteristics (solid line) for the DUT and and pulsed *I–V* characteristics (dash–dot line) for $v_{GS} = 0$ V. The ideal RF dynamic loadlines for different load impedances are shown for class-A operation (dotted lines). (From [24] with permission, ©2006 IEEE.)

4.3.1 CW dynamic loadline measurement system

CW dynamic loadlines and pulsed *I–V* characteristics for various DC-bias drain voltages are compared in Figure 4.10. As shown in Figure 4.10, the CW-RF loadlines measured for CW-RF signals at 2 GHz suffer from *I–V* knee walk-out. As the DC drain bias V_{DS} increases, the CW-RF loadlines fail to reach I_{dss}. As predicted in [6], they reach instead the pulsed *I–V* characteristics at v_{GS} of 0 V obtained using the same quiescent bias points (V_{DS} and $V_{GS} = -2.5$ V) as in the CW-RF loadlines.

4.3.2 Pulsed *I–V* pulsed-RF loadline measurement system

We have seen in the previous section that the thermal transient response is typically longer than a microsecond; the higher the thermal conductivity the slower the transient response. Since low-frequency memory effects have a slow time response, fast pulsed *I–V* pulsed-RF small-signal measurements will maintain the device temperature and traps at their quiescent state [25] [26]. This should hold also for large-signal pulsed-RF excitations, potentially allowing improved isothermal and iso-trapping RF device response in pulsed-bias operation if the device suffers from memory effects [24].

Both gate and drain pulses are now applied to the DUT with a pulse width of 1 μs, and a duty rate of 1%. The pulsed-RF signal, which is synchronized with the drain pulse, has a duration of 0.33 μs and duty rate of 0.33%. RF dynamic loadlines are then

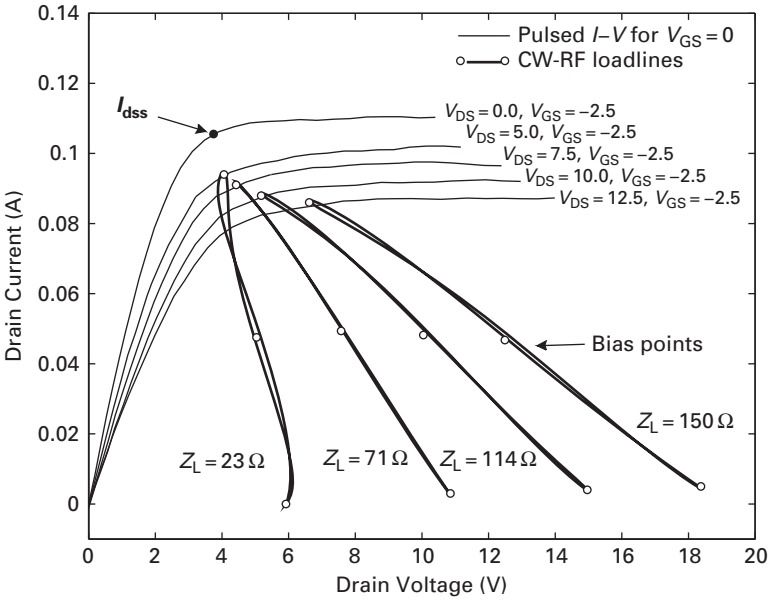


Figure 4.10 Dynamic loadline CW measurements at a constant substrate temperature T_{sub} of 30 °C. The 2-GHz loadlines are superposed on the $v_{GS} = 0$ V pulsed *I–V* characteristics, which are pulsed from the specified V_{DS} and V_{GS}.

reconstructed inside the pulse on the basis of frequency-domain data measured up to the fourth harmonics by the LSNA. It was verified that the pulsed-RF measurements give the same RF dynamic loadlines as those obtained by CW-RF measurements when constant DC biasing is used in both cases.

Figure 4.11 compares the RF dynamic loadlines obtained from CW-RF and pulsed *I–V* pulsed-RF measurements. In contrast to the CW-RF loadlines featuring a reduced current swing, the pulsed *I–V* pulsed-RF dynamic loadlines exhibit a much larger current swing. Using the current swing as a metric, the *I–V* knee walk-out is actually effectively suppressed. This is mainly due to the fact that the pulsed *I–V* pulsed-RF loadline measurements bypass the thermal and trap effects. Note that the negative currents in the pulsed *I–V* pulsed-RF dynamic loadlines result from both the energy storage in the drain capacitance and the self-biasing associated with the device nonlinearity.

It should be noted that the RF input power is precisely controlled to achieve the same gate-voltage swings of 5.0 V such that the device operates in all cases with a gate voltage ranging from −5 V, off state, to 0 V as shown in Figure 4.11(b).

4.3.3 Origin of the *I–V* knee walk-out in the CW-RF loadlines

Since the pulsed *I–V* pulsed-RF dynamic loadlines are free from the thermal and trap effects, it is natural to infer that the *I–V* knee walk-out in the CW-RF dynamic loadlines originates from the thermal and/or trap effects. By controlling the substrate temperature

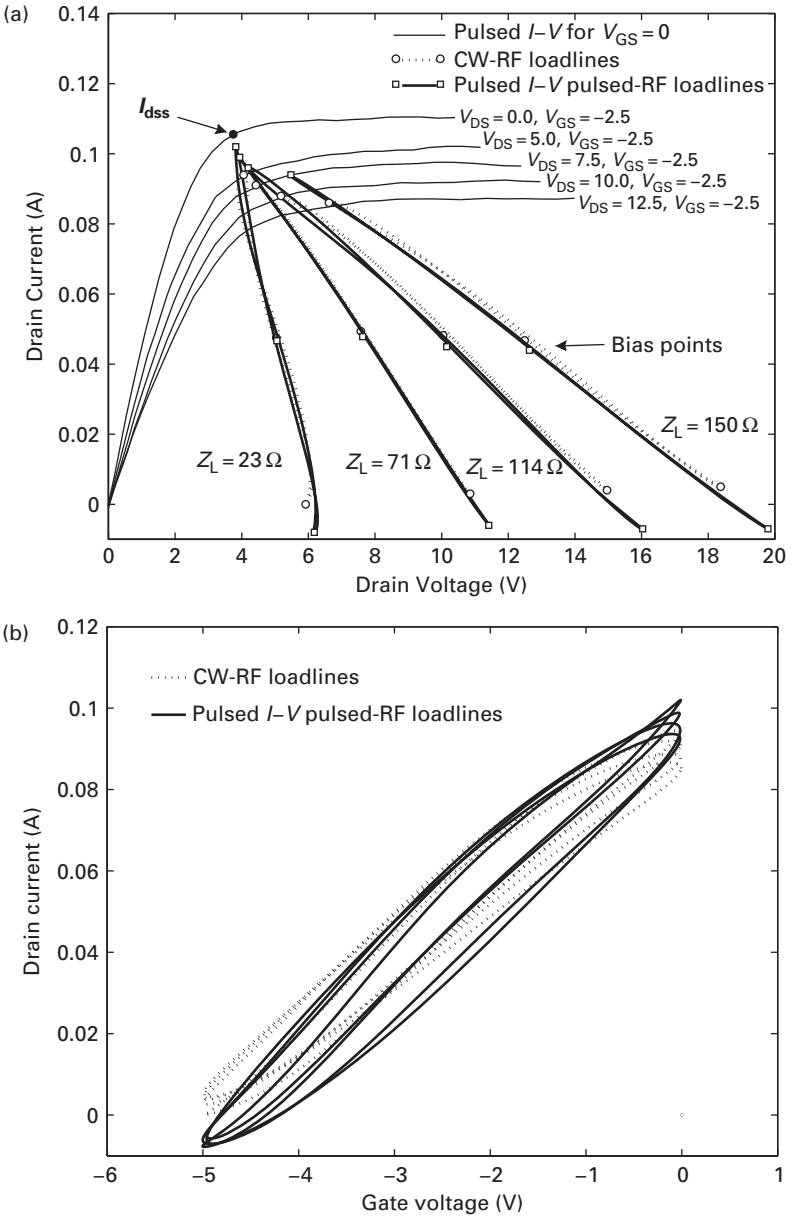


Figure 4.11 Comparison of the dynamic loadlines between CW-RF measurements (dots) and pulsed *I–V* pulsed-RF measurements (solid lines) at a constant substrate temperature T_{sub} of 30 °C. (a) RF dynamic drain loadlines. The loadlines are superposed on the $v_{GS} = 0$ V pulsed *I–V* characteristics which are pulsed from the specified V_{DS} and V_{GS}. (b) Gate loadlines. The quiescent voltage is $V_{GS} = -2.5$ V. (From [24] with permission, ©2006 IEEE.)

in the pulsed *I–V* pulsed-RF measurements, we shall now attempt to differentiate between the thermal and trap contributions.

Let us first determine the value of the device temperature resulting for a substrate temperature $T_{\text{sub,DC}}$ when the device is biased with V_{GS} and V_{DS}. The DC drain current observed at the bias points of interest can be expressed in terms of an isothermal current–voltage relation [1]:

$$I_{\text{D,DC}} = I_{\text{D,iso}}(V_{\text{GS}}, V_{\text{DS}}, T_{\text{dev,DC}}),$$

using the device temperature

$$T_{\text{dev,DC}} = T_{\text{sub,DC}} + R_{\text{th}}(T_{\text{sub,DC}}) \times P_{\text{avg}}$$

with R_{th} the effective thermal resistance and P_{avg} the average power dissipated by the device. The drain current under pulsed biasing from the off-state quiescent point ($V_{\text{GS}} = -5$ V and $V_{\text{DS}} = 0$ V) is given by

$$i_{\text{D,Pulsed}} = I_{\text{D,Pulsed}}(v_{\text{GS}}, v_{\text{DS}}, -5, 0, T_{\text{dev,Pulsed}}).$$

The self-heating effect which is present in CW operation can now be reproduced in the pulsed *I–V* biasing measurements by increasing the substrate temperature ($T_{\text{sub,Pulsed}}$) until the pulsed bias drain current ($I_{\text{D,Pulsed}}$) equals the DC bias drain current ($I_{\text{D,DC}}$), for $v_{\text{DS}} = V_{\text{DS}}$ and $v_{\text{GS}} = V_{\text{GS}}$,

$$i_{\text{D,Pulsed}}(V_{\text{GS}}, V_{\text{DS}}, -5, 0, T_{\text{sub,Pulsed}}) = I_{\text{D,DC}},$$

such that we have $T_{\text{sub,Pulsed}} = T_{\text{dev,Pulsed}} = T_{\text{dev,DC}}$. The resulting device temperatures $T_{\text{dev,DC}}$ are found to be 58, 72, 100, and 112 °C for V_{DS} of 5.0, 7.5, 10.0, and 12.5 V, respectively. The R_{th} of the DUT is then determined to be 130 °C/W in the wafer measurement setup.

The CW temperatures obtained from the pulsed *I–V* measurements are now also used in the pulsed *I–V* pulsed-RF measurement of the RF dynamic loadlines as shown in Figure 4.12. The comparison of the RF dynamic loadlines obtained from the CW-RF and the pulsed *I–V* pulsed-RF measurements under the above temperature of operation should then yield the trap contribution, since this should be the only remaining difference between these two measurements. However, the two loadlines are fairly similar for this passivated device. Table 4.3 shows the ratio of $\Delta I / I_{\text{dss}}$, where ΔI is the RF current swing and I_{dss} (0.105 A in Figure 4.11) is the maximum current at $v_{\text{GS}} = 0$ V which can be achieved from the knee region when the DUT is pulsed from the off state. As V_{DS} increases, it is observed that the RF dynamic loadlines of the pulsed *I–V* pulsed-RF measurements at the device temperature $T_{\text{dev,DC}}$ are indeed very similar to those of the CW-RF measurements at high bias.

This implies that the *I–V* knee walk-out at high bias results mostly from thermal effects rather than trap effects in this device. Table 4.3 also shows that the DUT provides around 100% RF performance for all terminations for pulsed *I–V* biasing at room

Table 4.3. Comparison of $\Delta I / I_{dss}$ between CW-RF and pulsed *I–V* pulsed-RF measurements at different temperatures

	V_{DS} (V)			
	5.0	7.5	10.0	12.5
CW-RF ($T_{dev,DC}$) (%)	91	85	81	78
Pulsed *I–V* pulsed-RF ($T_{dev,DC}$) (%)	98	91	85	81
Pulsed *I–V* pulsed-RF (30 °C) (%)	105	100	98	96

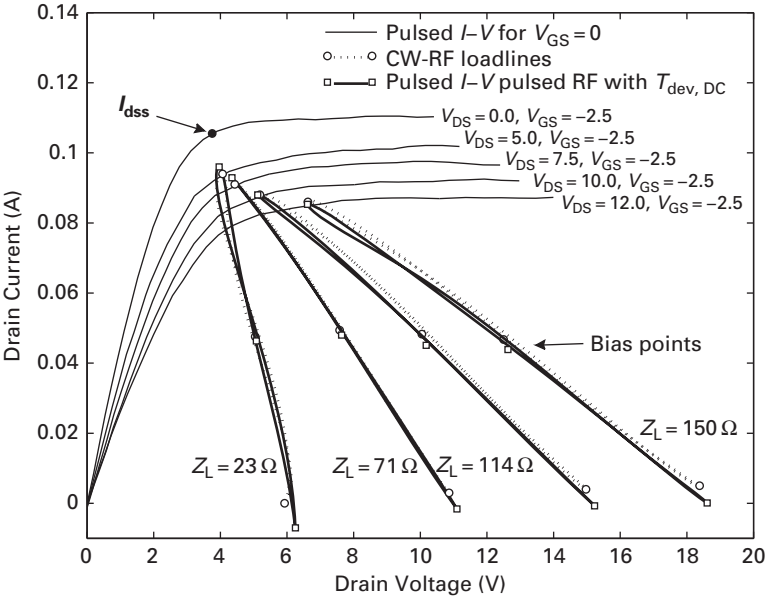


Figure 4.12 Dynamic loadlines (square and bold plain lines) for pulsed *I–V* pulsed-RF measurements with the substrate set to the normal CW device temperature $T_{sub} = T_{dev,DC}$ of 58, 72, 100, and 112 °C for V_{DS} of 5.0, 7.5, 10.0, and 12.5 V, respectively [27]. The CW loadlines (dashed lines) are also included for reference.

temperature. Note that the 5% excess current swing for V_{DS} of 5.0 V originates in part from the increase of the forward gain at 2 GHz due to the large series feedback of the DUT source resistance.

The thermal origin of the knee walk-out demonstrated in the above GaN HEMT is due to the use of sapphire as a substrate. As we shall see for an unpassivated GaN HEMT fabricated on SiC substrate, the knee walk-out originates instead from traps. However, some residual self-heating still needs to be accounted for.

4.4 Trapping in GaN HEMTs

Beside self-heating effects, trapping is a major source of memory effects in some devices. In this section we shall consider passivated and unpassivated GaN HEMTs on

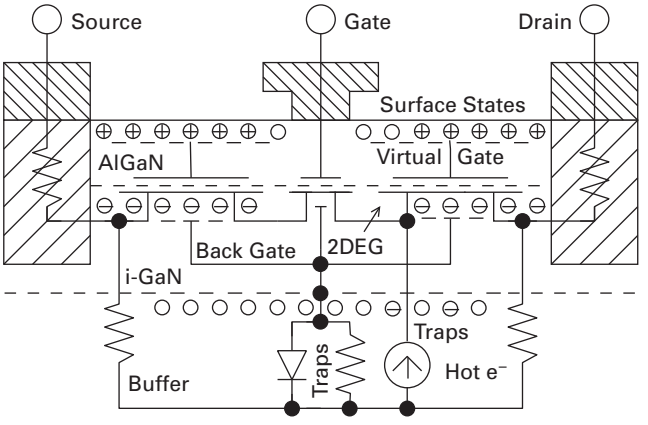


Figure 4.13 Layout and equivalent circuit for AlGaN surface trapping and GaN bulk trapping [28]. (From [29] with permission, ©2010 IEEE.)

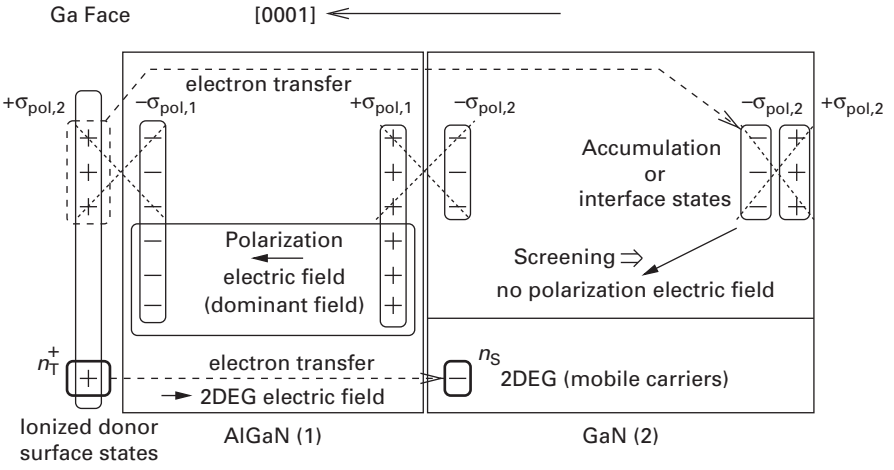


Figure 4.14 Proposed one-trap-level charge model in which the ionized donors provided the electron charges for both the 2DEG and the screening charges in the wide GaN region. Not shown are the additional piezoelectric charges induced by the stress in the pseudomorphic AlGaN material which add to the spontaneous polarization in the AlGaN region.

SiC substrate. Figure 4.13 shows various trapping mechanisms potentially taking place in these devices. They include (1) trapping at the AlGaN surface and (2) trapping in the buffer [28], which affect, respectively, the source and drain resistance and the threshold voltage.

The surface states on top of the AlGaN layer are believed to act as donors that sustain the two-dimensional electron gas (2DEG) in the gate–drain and gate–source regions [30]. A possible charge-balanced mechanism accounting for both the 2DEG and also the screening in the wide GaN region is shown in Figure 4.14. As the gate potential

relative to the source and drain potential becomes more negative, the electron population of the AlGaN surface donor traps increases (reduced ionization), which induces in turn a depletion of the 2DEG concentration. This is the so-called virtual gate effect [31]. The resulting increase in the source and drain parasitic resistances at high bias can then profoundly affect the DC and RF performance of the GaN HEMT devices [32]. In the subsequent sections, we present various RF measurements realized with an LSNA to investigate these memory effects.

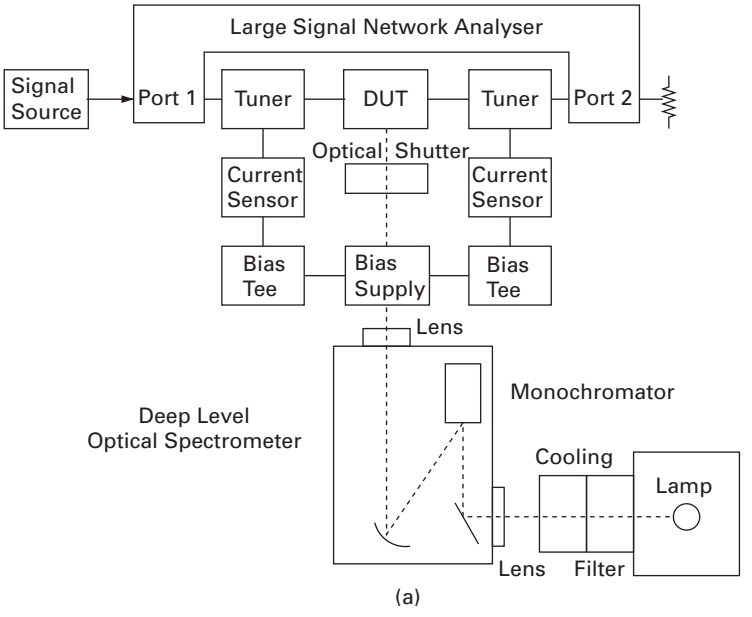
4.5 Characterization with a combined LSNA/DLOS system

The illumination of a microwave device with a monochromatic photon beam provides an additional probe for studying the trapping processes affecting its RF and DC operation. There are several physical contributions induced by illumination. First, by absorbing a photon of sufficiently large energy, a trapped electron can be emitted to the conduction band from a full trap level or a hole can be emitted in the valance band from an empty trap level. Second, an electron–hole pair can be generated once the photon energy exceeds the bandgap of AlGaN or GaN. The generated electron can directly contribute to the free electrons in the conduction band and the generated hole to free holes in the valence band. Hence, the interaction between illumination and trapped electrons is determined by the photon energy of the illumination and the energy level of the trapping states. These processes take place in addition to the normal thermal processes of emission and capture of electrons and holes from the conduction and valance bands, respectively.

The experimental setup of an LSNA/DLOS combined system [33] is depicted in Figure 4.15. Port 1 is used for the gate and port 2 for the drain of the device under test (DUT). The quiescent operating point of the DUT is set by the two external bias tees and DC power supplies. The voltage and current sensors detect the low-frequency time variations of the device bias voltage/current which are displayed on an oscilloscope. In the DLOS system, a 1000-W xenon lamp is used as a light source. The photon energy of the incident light tunable in the range 1.5–4.2 eV is generated through a monochromator. A shutter is placed between the focusing optics and the DUT to control the duration of the illumination.

Large-signal microwave measurements were performed under CW and pulsed monochromatic illumination conditions. The monochromatic illumination in unpassivated devices affects the large-signal RF response as demonstrated by the increase of the RF current and voltage swings in the device loadline in Figure 4.16.

Figure 4.17 shows the variation in drain-current swing of the RF loadline relative to the dark case as a function of the photon energy for both unpassivated and SiN-passivated HEMTs. The 4.0-eV level corresponds to the AlGaN bandgag (4.05 eV) and the 3.4-eV level corresponds to the GaN bandgap (3.44 eV). A sub-bandgap level is detected at 2.2 eV. This energy level is below the half bandgap of AlGaN. Time constants of 2 and 10 s are found to provide a good fit to the relaxation in



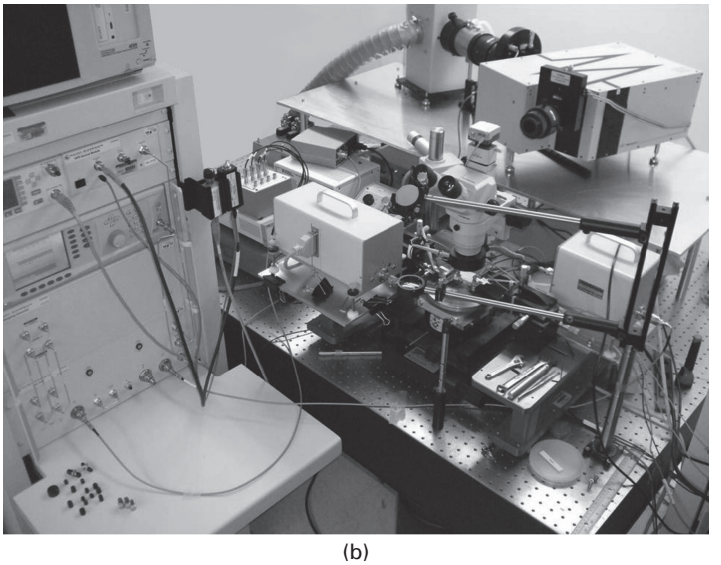


Figure 4.15 Experimental setup of LSNA–DLOS combined system: (a) a schematic diagram of the system and (b) and its implementation at Ohio State University.

time of the current swing when the illumination is turned off for 3.10 eV photon energy [33]. It is to be noted that the passivated device is less sensitive to the illumination. This desensitization can be interpreted, in view of the follow-up analysis, by invoking the fact that the drain resistance is already very low in the passivated device and cannot be decreased much by illumination (unlike in the unpassivated

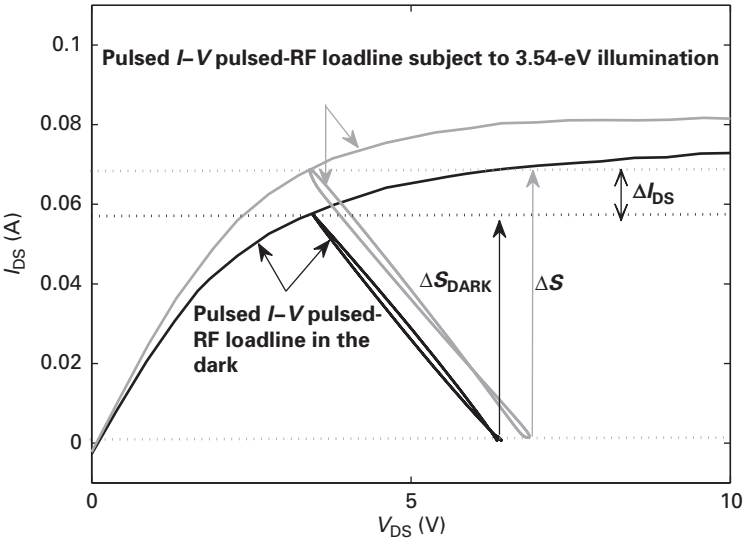


Figure 4.16 Large-signal RF loadline at 2 GHz measured under monochromatic illumination and in the dark. (From [33] with permission, ©2009 IEEE.)

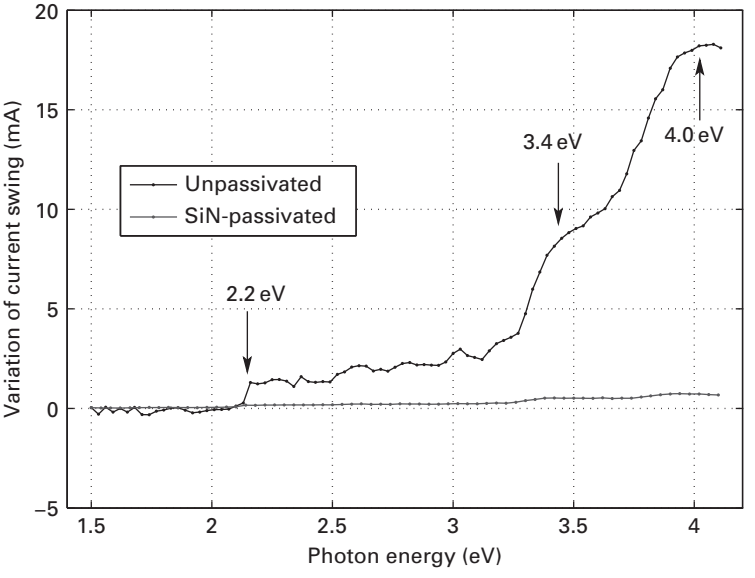


Figure 4.17 Photoionization spectrum of the variation of current swing for SiN-passivated and unpassivated AlGaN/GaN HEMTs. (From [33] with permission, ©2009 IEEE.)

device). Note also that the threshold voltage extracted from the dynamic loadline $I_D(t) - V_{GS}(t)$ exhibits a relatively small shift with illumination in the unpassivated devices investigated, suggesting that bulk traps are not the dominant mechanism in these devices.

4.6 Quasi-static device parasitics

On the basis of the RF characterization results in the previous section, it was found that the device is sensitive to illumination, suggesting trap activity. Similar results are also found using S-parameter measurements under illumination [33]. The magnitude of S_{21} is found to increase with illumination from low to high photon energies. The real part of S_{22} is found to decrease from 27 Ω to 15 Ω from low to high photon energies in the unpassivated device whereas it remains approximately constant in the passivated device. This suggests that the drain resistance R_D varies under illumination, since this is one of the dominant device parasitics located in the access region. Hence, it is reasonable to assume a relationship between the device performance and the device parasitics since the parasitics are known to have an important influence on the device characteristics.

The parasitics under various biases (V_{GS}, V_{DS}) can be extracted with the cold FET technique ($v_{DS} = 0$) by performing pulsed *I–V* pulsed-RF measurement at low duty rate [29]. The measurement setup used was presented in Section 2.9. The acquisition is done using multiple recording. The cold FET measurement proceeds as follows: (1) a quiescent bias point is selected, (2) the voltages of the gate and drain pulsed bias are tuned to achieve $v_{DS} = 0$ and $v_{GS} = 0$ during the short pulse duration, and (3) the pulsed RF signal is applied during that period for measuring the corresponding S-parameters. A nonlinear least-squares fitting method is utilized for fitting the measured S-parameters versus frequency from 0.7 to 13.2 GHz using the wave-equation-based equivalent circuit reported in [1] [29]. The same measurement procedure is used to determine the parasitic capacitances except that the transient bias conditions are now $v_{DS} = 0$ and $v_{GS} < V_T$.

Table 4.4 lists the extracted parasitic source and drain resistances measured at the quiescent bias of $V_{GS} = 0$ V and $V_{DS} = 0$ V. A deviation of the source and drain resistances from the reference values at $V_{DS} = 0$ and $V_{GS} = 0$ is observed when a different bias point is selected. Comparisons of the deviations of drain resistances for unpassivated and passivated devices are shown in Figure 4.18.

The deviation in source/drain resistance for the passivated device can be approximately described by the following equation:

$$\Delta R_{S/D} = R_{S/D0}\left(\frac{-\alpha \Delta T/T_0}{1+\alpha \Delta T/T_0}\right),$$

where $R_{S/D0}$ is the reference parasitic resistance, T_0 the reference temperature for μ_0, T the channel temperature, α the power of the temperature dependence, and ΔT the temperature deviation.

For the unpassivated device, the parasitics can be modeled as composed of two contributions: (1) near the gate contact, where both the surface states and the thermal effect influence the variation of the resistance, and (2) far from the gate contact, where the thermal effect dominates as in the passivated counterpart. Hence the deviation of the parasitic resistance can be expressed as

$$\Delta R_{S/D} = \frac{R_{S/D0}}{1+\alpha \Delta T/T_0}\left[-\alpha \frac{\Delta T}{T_0} - \frac{\Delta \ell_{S/D}}{\ell_{S/D}}\left(1 - \frac{1}{1+\beta_{S/D}}\right)\right],$$

Table 4.4. Reference parasitic resistances at the source and drain sides measured at the quiescent bias of $V_{GS} = 0$ V and $V_{DS} = 0$ V

	Unpassivated DUT	Passivated DUT
Source side R_{S0} (Ω)	6.27	5.77
Drain side R_{D0} (Ω)	22.36	22.88

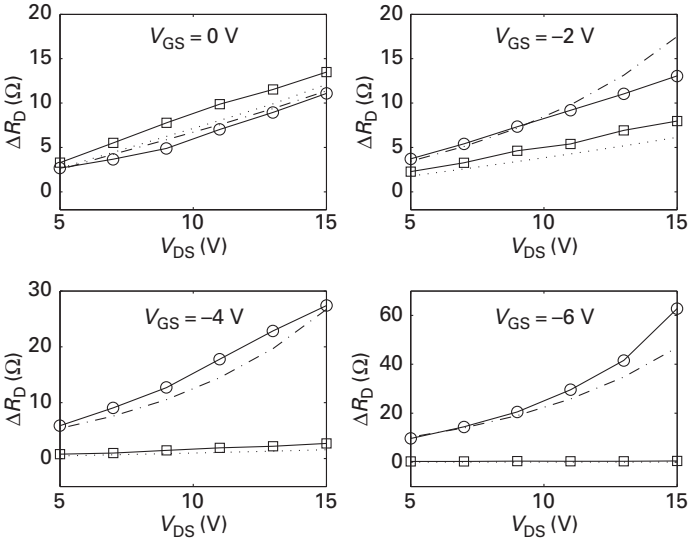


Figure 4.18 Deviation of drain resistance $\Delta R_D = R_D - R_{D0}$: measured (circle line for the unpassivated device and square line for the passivated device) and fitting (dash–dot line for the unpassivated device and dotted line for the passivated device) deviations of parasitic resistances on the drain side. (From [29] with permission, ©2010 IEEE.)

where $\ell_{S/D}$ is the length of the gate-to-source/drain spacer and $\Delta\ell_{S/D}$ the width of the depleted region in the spacers. The factor $\beta_{S/D} = (n_s - n_{s,0})/n_{s,0}$ is the modulation index of the 2DEG concentration in the depleted region, which varies between 0 and −1. Empirical functional dependences are introduced for fitting the bias dependences of β and $\Delta\ell/\ell$ both on the source side and on the drain side:

$$\beta = \frac{n_s - n_{s,0}}{n_{s,0}} = \tanh(C_1 V_{GS} + C_2 V_{GD}),$$
$$\frac{\Delta\ell}{\ell} = C_3 \tanh(C_4 V_{GS}),$$

where C_1, C_2, C_3, and C_4 are fitting parameters. Theoretical justification for these expressions will be derived in the next section. A typical plot of the device I–V characteristics together with a contour plot of the 2DEG concentration modulation index β is shown in Figure 4.19 for an unpassivated device.

The change of the parasitic resistance ΔR_D with the gate bias is also accompanied by a change of the channel temperature ΔT, which is plotted in Figure 4.20 for the

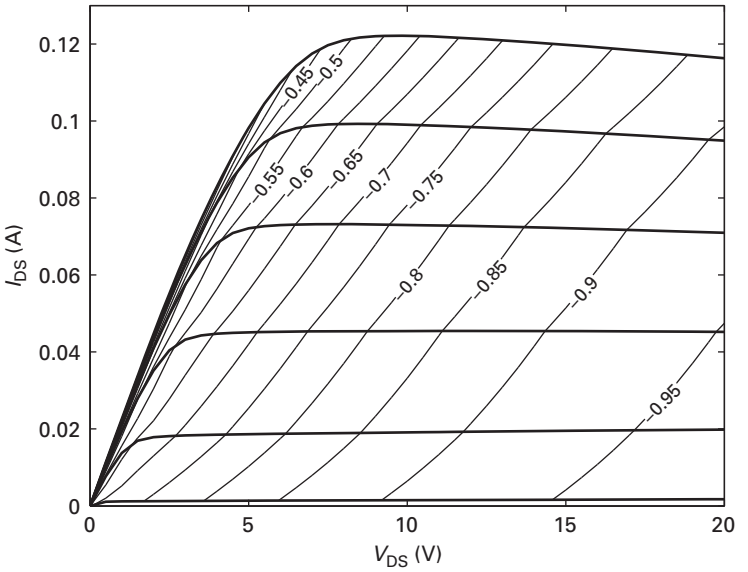


Figure 4.19 An I–V characteristic plot and contour plot of the 2DEG concentration modulation index β for an unpassivated GaN HEMT. Walk-out is canceled for $\beta = 0$. (From [33] with permission, ©2010 IEEE.)

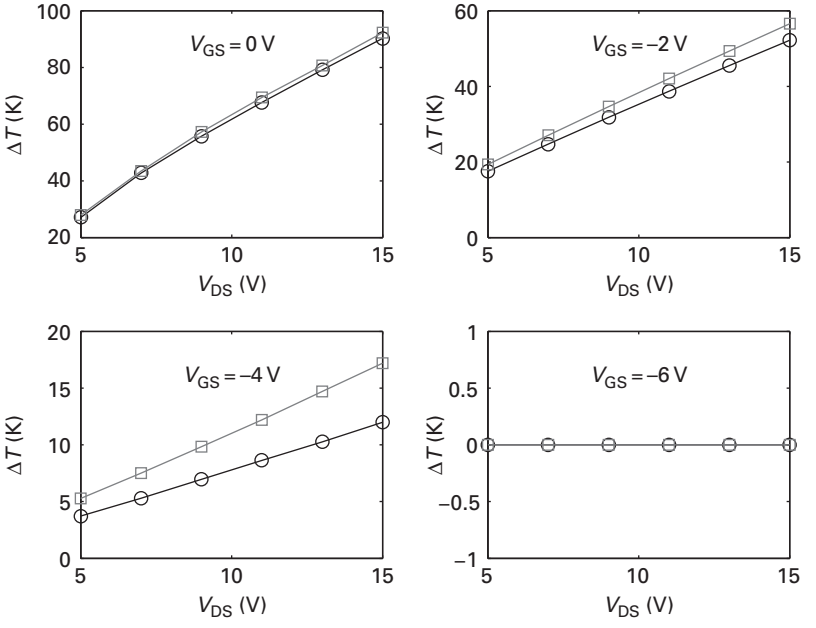


Figure 4.20 Calculated change of the channel temperature with various quiescent bias points for the unpassivated (circles) and passivated (squares) devices. (From [29] with permission, ©2010 IEEE.)

unpassivated (squares) and passivated (circles) devices. A large self-heating is predicted for both types of devices.

4.7 Rate equation for physical modeling of trapping effects

In the previous section a phenomenological model was introduced to model the dependence of the 2DEG channel modulation index β upon the gate and drain voltages. A simple non-equilibrium model is presented now to bring more insights into the physical process taking place.

In the absence of illumination, when the device is biased with non-zero V_{GS} and V_{DS} (non-equilibrium), a rate equation for a single-energy electron trap level located at the AlGaN surface can be defined using the quasi-Fermi level of the trap $E_{F,T}(V_{GS}, V_{DS})$ as follows:

$$\begin{aligned}
\frac{dn_T}{dt} &= \frac{N_T}{\tau_c} - \frac{n_T}{\tau}, \\
\frac{1}{\tau} &= \frac{1}{\tau_c} + \frac{1}{\tau_e}, \\
n_T &= N_T f_T, \\
f_T &= \frac{1}{1 + \exp\left[(E_T - E_{f,T})/(k_B T)\right]}, \\
\frac{\tau_c}{\tau_e} &= \frac{1 - f_T}{f_T} \quad \text{in steady-state} \quad \left(\frac{dn_T}{dt} = 0\right),
\end{aligned}$$

where N_T is the total trap concentration, n_T is the concentration of occupied trap levels, $\tau_c = \tau_{c,th}$ is the capture time, $\tau_e = \tau_{e,th}$ is the emission time, E_T is the trap energy level, k_B is the Boltzmann constant, and T is the absolute temperature.

In AlGaN/GaN HEMT structures, the existence of donor-like surface states has been proposed as the principal source of the formation of the 2DEG [30]. It is assumed that the unoccupied (ionized) donor-like electron traps with a concentration are imaging the 2DEG with concentration n_s in the conducting channel of the AlGaN/GaN HEMT:

$$n_T^+ = N_T - n_T = (1 - f_T)N_T = n_s.$$

For a single-trap model, this holds in steady state for any bias condition with or without illumination. In equilibrium, when $V_{GS} = 0$ V and $V_{DS} = 0$ V, and no illumination is applied, n_s relaxes to the equilibrium 2DEG concentration $n_{s,0}$ and the quasi-Fermi level at the traps is the same as the 2DEG Fermi level $E_{F,T} = E_{F,0}$ as shown in Figure 4.21. Under such equilibrium conditions, the capture/emission time $\tau_{c/e}$ is the thermal capture/emission time $\tau_{c/e} = \tau_{c/e,th}$, satisfying

$$\frac{\tau_e}{\tau_c} = \frac{\tau_{e,th}}{\tau_{c,th}} = \exp\left(\frac{E_{F,T} - E_T}{k_B T}\right).$$

When a bias is applied, the traps are no longer in equilibrium with the 2DEG due to the injection of electrons from the surface. Under such non-equilibrium conditions, the

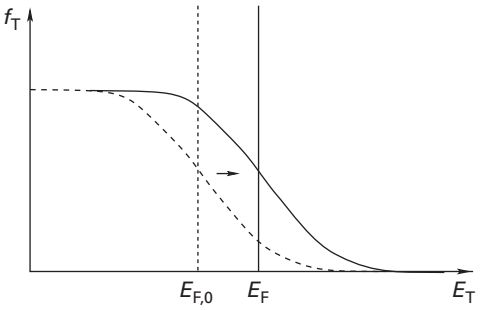


Figure 4.21 f_T distribution at temperature T as a function of the trap energy level E_T. $E_{F,T} = E_{F,0}$ is the device Fermi level when no DC bias is applied. The trap quasi-Fermi level moves to $E_{F,T} = E_F$ when a bias is applied to the AlGaN/GaN device due to the injection of electrons from the AlGaN surface. E_T represents the single-trap energy level, which can be either lower or higher than the quasi-Fermi level.

capture time τ_c is reduced due to the mobility of the electrons at the AlGaN surface of the unpassivated device:

$$\frac{1}{\tau_c} = \frac{1}{\tau_{c,th}} + \frac{1}{\tau_{c,tunnel}},$$

where $\tau_{c,tunnel}$ is the surface tunneling/hopping time. This process is assumed to be essentially non-reversible ($\tau_e \simeq \tau_{e,th}$) due to the Fermi exclusion factor $(1 - f_{T,gate})$ on the gate supply side. The trap population can then be described using a bias-dependent quasi-Fermi level for the traps $E_{F,T}(V_{GS}, V_{DS})$ as shown in Figure 4.21. The relative position between the quasi-Fermi level and the trap energy level E_T then determines the corresponding change in the 2DEG charge n_s when a bias is applied. It results that we have the following relation in steady state:

$$\frac{n_{s,0}}{N_T} = 1 - f_T(E_{F,0}), \tag{4.8}$$

$$\frac{n_s}{N_T} = 1 - f_T(E_{F,T}), \tag{4.9}$$

and the ratio of the capture to emission time $\gamma = \tau_c/\tau_e$ for the trap level is then given by the trap occupation f_T. From Equation 4.9, the deviation of n_s with respect to $n_{s,0}$ in steady state when a bias is applied can be expressed as

$$\begin{aligned}\beta = \frac{n_s - n_{s,0}}{n_{s,0}} &= \frac{f_T(E_{F,0}) - f_T(E_{F,T})}{1 - f_T(E_{F,0})}, \\ &= \frac{\exp\left(\dfrac{E_{F,0} - E_F}{2k_BT}\right) - \exp\left(\dfrac{E_F - E_{F,0}}{2k_BT}\right)}{\exp\left(\dfrac{E_T}{k_BT} - \dfrac{E_{F,0} + E_F}{2k_BT}\right) + \exp\left(\dfrac{E_F - E_{F,0}}{2k_BT}\right)}, \\ &= \frac{\exp(x) - \exp(-x)}{\exp\left[(E_T - E_{F,0})/(k_BT) + x\right] + \exp(-x)},\end{aligned} \tag{4.10}$$

using

$$x = \frac{E_{\mathrm{F},0} - E_{\mathrm{F}}}{2k_{\mathrm{B}}T}.$$

Note that $\beta = (n_{\mathrm{s}} - n_{\mathrm{s},0})/n_{\mathrm{s},0}$ varies from 0 to -1 as x varies from 0 (equilibrium) to $-\infty$.

In the previous section, we have introduced a phenomenological expression for the bias dependence of $\beta = (n_{\mathrm{s}} - n_{\mathrm{s},0})/n_{\mathrm{s},0}$, the modulation index of the 2DEG concentration:

$$\beta = \tanh(C_1 V_{\mathrm{GS}} + C_2 V_{\mathrm{GD}}).$$

Note that in the specific case of $E_{\mathrm{T}} = E_{\mathrm{F},0}$ (trap half occupied) Equation (4.10) yields $\beta = \tanh(x)$ and we can then identify x to be given by

$$x = \frac{E_{\mathrm{F},0} - E_{\mathrm{F}}}{2k_{\mathrm{B}}T} = C_1 V_{\mathrm{GS}} + C_2 V_{\mathrm{GD}}.$$

Thus, this simple non-equilibrium single-trap-level model which relies on the concept of the quasi-Fermi level brings some physical justification for the phenomenological expressions which were needed to fit the experimental data.

4.8 Two-trap-level model

In the previous section a model with a single deep donor level was presented. However, a very large emission and capture time constant (days) is usually associated with the deep energy levels, whereas faster processes have also been observed when the system relaxes toward equilibrium. More complex models can be developed to account for such processes.

The next level of complexity consists in a two-trap-level model involving a deep donor trap ($E_{\mathrm{c}} - 1.65$ eV in this example) and a shallower acceptor trap ($E_{\mathrm{c}} - 0.35$ eV in this example). The 2DEG charge is then supported by

$$n_{\mathrm{s}} = \big(1 - f_{\mathrm{T,D}}\big)N_{\mathrm{T,D}} - f_{\mathrm{T,A}}N_{\mathrm{T,A}}.$$

In equilibrium the shallower acceptor levels are not populated and the 2DEG is given by $n_{\mathrm{s}} \simeq \big(1 - f_{\mathrm{T,D}}[E_{\mathrm{F},0}]\big)\, N_{\mathrm{T,D}}$. The resulting equilibrium band diagram is shown in Figure 4.22 for $N_{\mathrm{T}} = 2 \times 10^{17}$ m^{-2}. The 2DEG charge concentration of 8.7×10^{16} m^{-2} is obtained with a self-consistent solution of the Schrödinger equation.

Under non-equilibrium conditions (V_{GD} and V_{GS} non-zero), electrons are transferred from the gate to the acceptor states while the donor states are assumed to remain frozen due to their long time constants. Enforcing current conservation from the electron moving on the AlGaN surface, transferring to the AlGaN conduction band via the acceptor traps at the surface, and moving across the entire AlGaN layer enables one to calculate the acceptor trap occupation $f_{\mathrm{T,A}}$ and therefore n_{s} as a function of the tunneling/hopping time constant $\tau_{\mathrm{c,tunnel}}$. The resulting band diagram under non-equilibrium conditions is shown in Figure 4.23. The quasi-Fermi level of the electrons

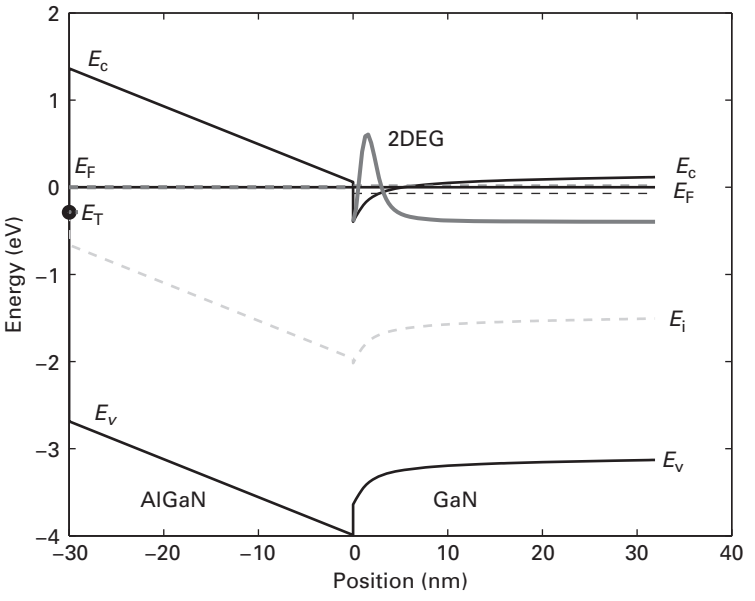


Figure 4.22 Equilibrium band diagram assuming a single-trap concentration of 2×10^{17} m^{-2}. The trap level is indicated by a black circle.

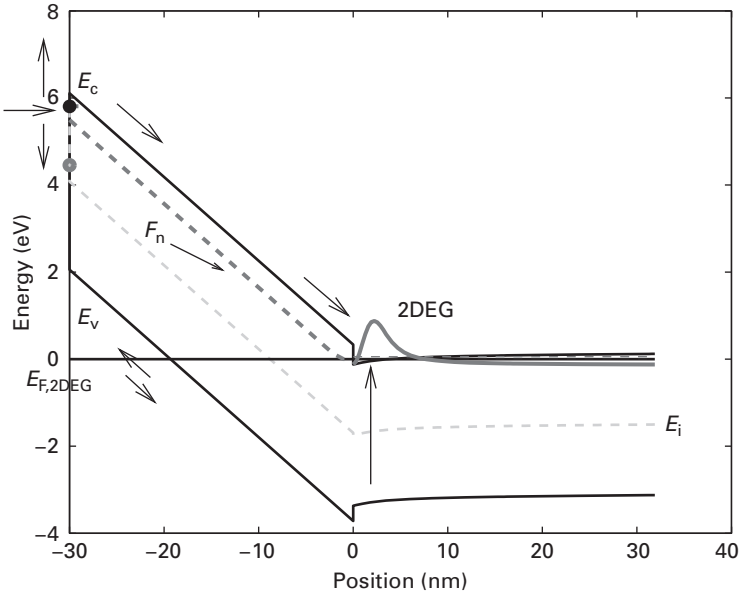


Figure 4.23 Band diagram for the two-trap-level model (1.65 eV and 0.35 eV below E_c) under non-equilibrium conditions. The deep energy level (gray circle) is a donor surface state. The shallower energy level (black circle) is an acceptor surface state.

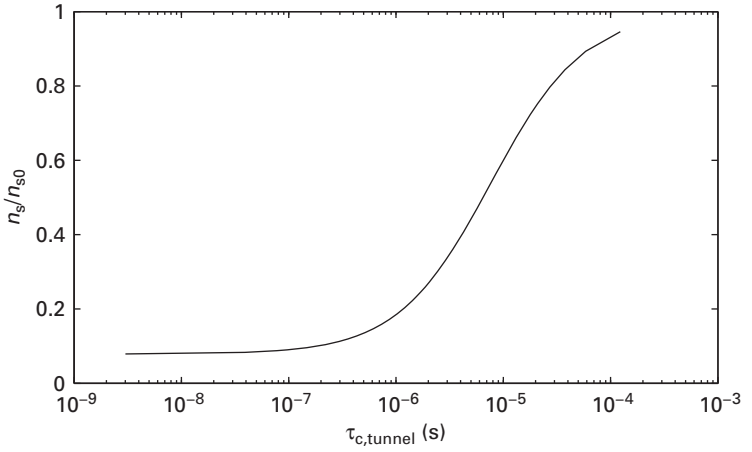


Figure 4.24 Normalized 2DEG concentration versus tunneling time constant.

F_n is now closer to the conduction band, inducing a larger non-equilibrium electron charge supporting the collection by the 2DEG of the electrons emitted by the trap. A larger tunneling current can then be handled in steady state by the shallower acceptor states than could be by the deeper donor states.

The resulting relation between the 2DEG concentration n_s and the tunneling time constant $\tau_{c,tunnel}$ is given in Figure 4.24. This two-trap-level model will be used in the next section to describe the cyclostationary effect which may be observed under large-signal RF excitations.

4.9 Cyclostationary effect

4.9.1 Theory

Emission and capture of electrons by traps typically have long time constants. Yet it has been observed that under large-signal high-frequency periodic excitation the average trap population responds to the high-frequency periodic signal applied, yielding in some cases a dramatic increase or decrease of the occupation and noise generated by the traps. This is the so-called cyclostationary effect [34].

Strong cyclostationary effects will normally be observed under periodic excitations when the emission and capture rate have a nonlinear dependence upon the instantaneous gate and drain voltages. Consider the following single-trap cyclostationary process where the emission $\tau_e(t)$ and capture $\tau_c(t)$ time constants are modulated by the RF signals [35]:

$$\frac{df_{T,A}(t)}{dt} + \left(\frac{1}{\tau_e(t)} + \frac{1}{\tau_c(t)}\right) f_{T,A} = \frac{1}{\tau_c(t)}.$$

It has been theoretically established [35] that, for the periodic (RF) excitation of such a system, the time average of the capture and emission rates should be used for evaluating the average trap population $\langle f_{\mathrm{T,A}}(t)\rangle$ and the noise spectral density $S_{W_{\mathrm{g}}\Delta\ell_{\mathrm{D}}n_{\mathrm{T,A}}}$ (burst or popcorn noise) for $n_{\mathrm{T,A}} = f_{\mathrm{T,A}}(t)N_{\mathrm{T,A}}$:

$$\langle f_{\mathrm{T,A}}\rangle = \frac{\langle g_{\mathrm{c}}\rangle}{\langle g_{\mathrm{c}}\rangle + \langle g_{\mathrm{e}}\rangle} = \frac{\langle g_{\mathrm{c}}\rangle}{\langle \lambda_0\rangle},$$
$$S_{W_{\mathrm{g}}\Delta\ell_{\mathrm{D}}n_{\mathrm{T,A}}} = N_{\mathrm{T,A}}\langle f_{\mathrm{T,A}}\rangle\big(1 - \langle f_{\mathrm{T,A}}\rangle\big)\frac{2\langle\lambda_0\rangle}{1 + \langle\lambda_0\rangle^2\,\Delta\omega^2}W_{\mathrm{g}}\,\Delta\ell_{\mathrm{D}}, \tag{4.11}$$

where we have $\langle\lambda_0\rangle = \langle g_{\mathrm{c}}\rangle + \langle g_{\mathrm{e}}\rangle$ using the definitions

$$\langle g_{\mathrm{c}}\rangle = \frac{1}{T}\int_0^T \frac{1}{\tau_{\mathrm{c}}[v_{\mathrm{GS}}(t), v_{\mathrm{DS}}(t)]}dt \qquad \text{and} \qquad \langle g_{\mathrm{e}}\rangle = \frac{1}{T}\int_0^T \frac{1}{\tau_{\mathrm{e}}[v_{\mathrm{GS}}(t), v_{\mathrm{DS}}(t)]}dt$$

with T the RF period, W_{g} the gate width, and $\Delta\ell_{\mathrm{D}}$ the width of the pinchoff region in the gate-to-drain spacer. Under large-signal excitations, the averages $\langle g_{\mathrm{c/e}}[v_{\mathrm{GS}}(t), v_{\mathrm{DS}}(t)]\rangle$ will differ from $g_{\mathrm{c/e}}[V_{\mathrm{GS,dc}}, V_{\mathrm{DS,dc}}]$, due to the nonlinear dependence of $g_{\mathrm{c/e}}$ upon v_{GS} and v_{DS}, thus yielding the cyclostationary effect.

In the two-trap-level models we have presented in the previous section, only the tunneling time constant can be physically modulated by the RF applied bias. Field-assisted hopping of the electron at the AlGaN surface is modulated by the RF voltage and so is consequently the tunneling time constant $\tau_{\mathrm{c,tunnel}}$ of the acceptor levels. On the other hand, we assume that the thermal capture rate, which is a function of the electron and hole concentrations in AlGaN, is negligibly modulated by the instantaneous RF gate-to-drain voltage due to the slow emission rate. Note that the thermal emission remains time-independent as long as the device temperature does not change. Thus the tunneling capture rate should be the dominant rate instantaneously modulated by the RF bias via the modulation of the gate-to-drain potential barrier at the AlGaN surface.

4.9.2 Experimental investigations

Using a source and a load tuner, the impact of the load resistance R_{L} on the device characteristic can be investigated to reveal the cyclostationary effect.

The measured RF drain loadlines of the unpassivated GaN HEMT are shown in Figure 4.25 for various external loads R_{L} from 50 to 170 Ω together with the DC (solid lines) and pulsed (dashed lines) *I–V* characteristics from $V_{\mathrm{GS}} = -3$ V, $V_{\mathrm{DS}} = 10$ V and a contour plot of the modulation index β in the drain spacer region next to the gate. The gate voltage in the *I–V* characteristics is swept from −5 V to 0 V in steps of 1 V.

In the previous chapter the modulation index $\beta = (n_{\mathrm{s}} - n_{\mathrm{s0}})/n_{\mathrm{s,0}}$ was defined as the deviation in channel concentration from the equilibrium channel concentration $n_{\mathrm{s,0}}$ [29]. Note that β varies from 0 to −1 as n_{s} varies from the equilibrium 2DEG concentration $n_{\mathrm{s,0}}$ to 0 (fully depleted 2DEG).

As shown in the contour plot in Figure 4.25, β decreases as both the gate voltage and the drain voltage deviate from 0 V. In Figure 4.25, circles represent the DC biasing points while squares indicate the effective DC operating points for the traps due to the cyclostationary process (see the detailed discussion below). It is observed that the

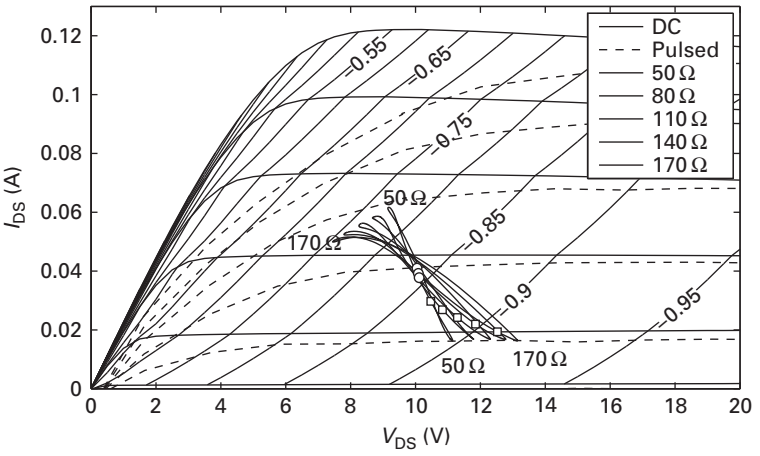


Figure 4.25 RF drain loadlines of the unpassivated GaN HEMT with DC (solid lines) and pulsed (dashed lines) *I–V* characteristics from $V_{GS} = -3$ V, $V_{DS} = 10$ V, and a contour plot of β. For both *I–V* characteristics, V_{GS} is swept from -5 V (bottom) to 0 V (top) in steps of 1 V. The solid lines, from top to bottom, are for external loads from 50 Ω to 170 Ω as indicated. (From [36] with permission, ©2009 IEEE.)

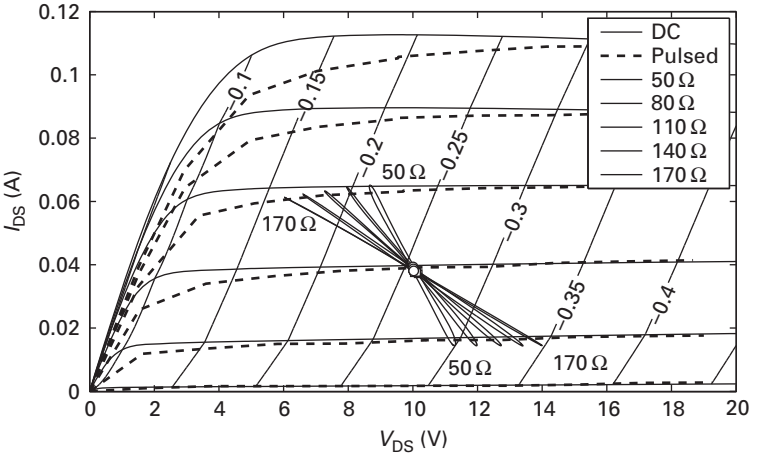


Figure 4.26 RF drain loadlines of the passivated GaN HEMT with DC (solid lines) and pulsed (dashed lines) *I–V* characteristics from $V_{GS} = -3$ V, $V_{DS} = 10$ V, and a contour plot of β. For both *I–V* characteristics, V_{GS} is swept from -5 V (bottom) to 0 V (top) in steps of 1 V. The solid lines, from top to bottom, are for external loads from 50 Ω to 170 Ω as indicated. (From [36] with permission, ©2009 IEEE.)

drain-voltage swing is approximately proportional to the load impedance. The loadlines are seen to bend at low drain voltages with increasing load resistance R_L. This is due to the increased knee walk-out induced by the shifting of the effective DC operating point of the traps.

Similar measurements with the passivated GaN HEMT are shown in Figure 4.26. β decreases now only to a minimum value of -0.26 such that there is no appreciable knee

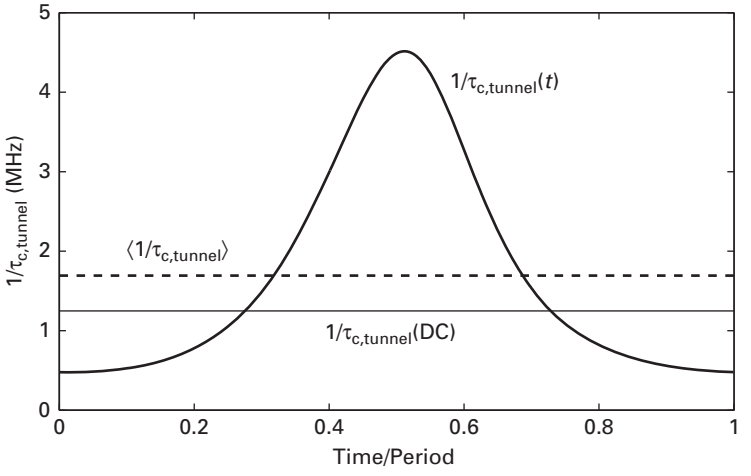


Figure 4.27 Variation of the tunneling capture rate $1/\tau_{c,tunnel}(t)$ versus time t over a period T for the unpassivated device.

walk-out. Thus, the loadlines are not distorted (straight lines) since the device remains operating in the saturation region. The overall range of β for passivated devices is higher than that of the unpassivated devices.

The variation of the tunneling capture rate (thick plain line) versus time for the unpassivated device is shown in Figure 4.27. The time-averaged capture rate $\langle 1/\tau_{c,tunnel}\rangle$ (dashed line) is seen to be larger than its DC value $1/\tau_{c,tunnel}(V_{GS,dc}, V_{DS,dc})$ (thin plain line) at the operating point for the unpassivated device. The nonlinear variation of $1/\tau_{c,tunnel}(t) = 1/\tau_{c,tunnel}[v_{GS}(t), v_{DS}(t)]$ versus time for the unpassivated device originates from the rapid decrease in Figure 4.24 of the tunneling capture time $\tau_{c,tunnel}(n_s)$ for low 2DEG concentrations. Note that the relation $\tau_{c,tunnel}(n_s)$ in Figure 4.24 and the bias dependence of $\beta(v_{GS}, v_{DS})$ are used to establish the dependence of the tunneling capture time $\tau_{c,tunnel}$ upon the bias voltage:

$$\tau_{c,tunnel}(n_s) = \tau_{c,tunnel}\,[n_{s0}\{1 + \beta(v_{GS}, v_{DS})\}].$$

The increase of the time-averaged capture rate $\langle 1/\tau_{c,tunnel}\rangle$ over its DC value augments the population of the shallow acceptor surface states, yielding a reduced 2DEG concentration. In the slower one-trap-level model this would be equivalent to the AlGaN surface current providing charges compensating the donor surface states.

The effective trap operating points for each of the loadlines (shown in Figures 4.25 and 4.26 using squares) correspond to the smallest drain voltage on the loadline which would generate the same average acceptor trap occupation (same β value) under DC conditions (no RF applied). These effective DC operating points for the traps indicate an average β ranging from -0.86 to -0.91 for R_L varying from 50 Ω to 170 Ω. Clearly the cyclostationary effect modifies the DC trap occupation, yielding in turn increased degradation in the *I–V* characteristics compared with the pulsed *I–V* characteristics from the DC bias points. The increased *I–V* knee walk-out in turn degrades the RF response of the device, as indicated by the distorted loadlines measured for the unpassivated device

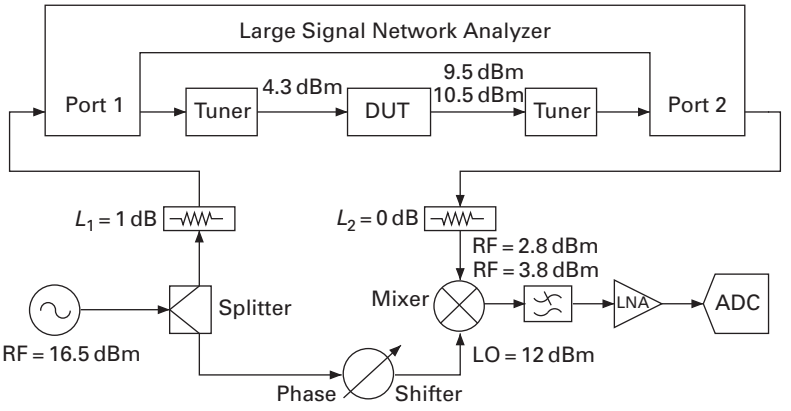


Figure 4.28 Block diagram of the additive phase noise measurement setup [37] integrated with an LSNA. The power levels throughout the measurement system for both unpassivated and passivated devices are also indicated.

in Figure 4.25. In contrast, for the passivated device in Figure 4.26, the dynamic RF loadlines are in good agreement with the pulsed *I–V* characteristics due to the negligible fluctuation in trapping at the AlGaN surface.

It also interesting to measure the RF noise presumably induced by the trapping activity. The setup used for the additive noise measurement [37] is shown in Figure 4.28. This setup is integrated with the NVNA such that the additive noise can be measured at the same time as the loadlines are acquired. With a variation of the phase shifter by 90°, either the additive amplitude (AM) or the additive phase noise (PM) can be measured in the RF signal [37].

The additive phase noise results obtained as a function of the output impedance termination are shown in Figure 4.29 for both unpassivated and passivated devices. The additive phase noise shown in Figure 4.29 does not exhibit a Lorentzian dependence as in Equation (4.11) but varies instead approximately as $1/\Delta f$ (with Δf the RF offset frequency) for both passivated and unpassivated devices. The additive amplitude noise is also found to vary as $1/\Delta f$ ($1/f$ noise) in the unpassivated device but is typically 10 dB higher than the phase noise. This is to be expected since the fluctuating excess drain resistance ΔR_{D} acts as a gain modulator in the unpassivated device. The occurrence of $1/f$ noise is usually assumed to be indicative of the presence of a network of traps having a continuum of energy levels or distributed in space over a continuum of distances from the surface/interface [38]; but other mechanisms have been proposed [39].

As is shown in Figure 4.29, the additive RF phase noise measured in the drain current noise is found to increase with increasing load resistance R_{L} in both passivated and unpassivated devices [36]. A conventional $1/f$-noise model [38] can indeed be developed for the unpassivated device. This links the increase in additive RF noise to the increase of the excess drain resistance ΔR_{D}^2. When the load resistance R_{L} increases, the additive noise will then increase as the acceptor trap occupation and ΔR_{D} increase as a consequence of the cyclostationary effect.

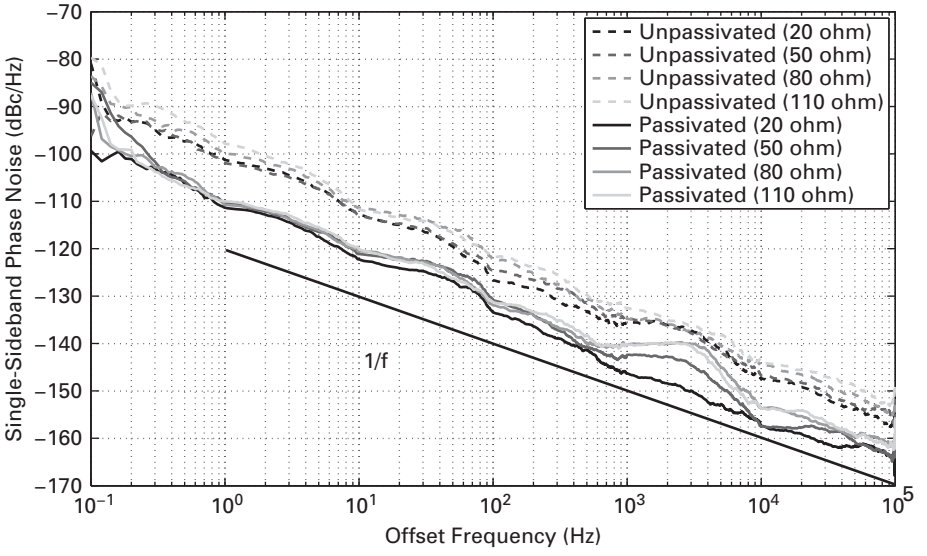


Figure 4.29 Comparison of phase-noise dependence between passivated and unpassivated devices on load impedances. (From [36] with permission, ©2009 IEEE.)

The additive phase RF drain-current noise for the passivated device is also found to be 10 dB smaller than for the unpassivated device as a result of the stabilization of the source and drain resistance by the passivation. The drain resistance is then no longer the dominant contribution for the $1/f$ noise observed in passivated devices since ΔR_D has vanished.

Thus, the proposed two-trap-level model for the AlGaN surface offers a plausible picture for the complex physical processes taking place in unpassivated AlGaN/GaN HEMTs. In passivated AlGaN/GaN HEMTs, thermal memory effects (mobility and saturation velocity degradation) and possibly bulk traps (threshold shift) are expected to play a more active role. The drain-to-source capacitance can also be responsible for inducing a large current drain lag and effective knee walk-out at high frequencies for large DC drain biases.

References

[1] P. Roblin and H. Rohdin, *High-Speed Heterostructure Devices*, Cambridge University Press, 2002.

[2] G. H. Jessen, R. C. Fitch, J. K. Gillespie, G. D. Via, N. A. Moser, M. J. Yannuzzi, A. Crespo, J. S. Sewell, R. W. Dettmer, T. J. Jenkins, R. F. Davis, J. Yang, M. A. Khan, and S. C. Binari, "High performance 0.14 μm gate-length AlGaN/GaN power HEMTs on SiC," *IEEE Electron Device Letters*, Vol. 24, No. 11, pp. 677–679, Nov. 2003.

[3] S. C. Binari, P. B. Klein, and T. E. Kazior, "Trapping effects in GaN and SiC microwave FETs," *Proceedings of the IEEE*, Vol. 90, No. 6, pp. 1048–1058, June 2002.

[4] R. Vetury, N. Q. Zhang, S. Keller, and U. K. Mishra, "The impact of surface states on the DC and RF characteristics of AlGaN/GaN HFETs," *IEEE Transactions on Electron Devices*, Vol. 48, No. 3, pp. 560–566, Mar. 2001.

[5] S. C. Binari, K. Ikossi, J. A. Roussos, W. Kruppa, D. Park, H. B. Dietrich, D. D. Koleske, A. E. Wickenden, and R. L. Henry, "Trapping effects and microwave power performance in AlGaN/GaN HEMTs," *IEEE Transactions on Electron Devices*, Vol. 48, No. 3, pp. 465–471, Mar. 2001.

[6] P. McGovern, J. Benedikt, P. J. Tasker, J. Powell, K. P. Hilton, J. L. Glasper, R. S. Balmer, T. Martin, and M. J. Uren, "Analysis of DC-RF dispersion in AlGaN/GaN HFETs using pulsed *I–V* and time-domain waveform measurements," in *IEEE MTT-S International Microwave Symposium Digest*, Long Beach, CA, 2005.

[7] Y.-F. Wu, A. Saxler, M. Moore, R. P. Smith, S. Sheppard, P. M. Chavarkar, T. Wisleder, U. K. Mishra, and P. Parikh, "30 W/mm GaN HEMTs by field plate optimization," *Electron Device Letters*, Vol. 25, No. 3, pp. 117–119, Mar. 2004.

[8] J. T. Hsu and L. V. Quoc, "A rational formulation of thermal circuit models for electro-thermal simulation – part I: finite element method," *IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications*, Vol. 43, No. 9, pp. 721–732, Sept. 1996.

[9] K. Fukahori and P. R. Gray, "Computer simulation of integrated circuits in the presence of electro-thermal interaction," *IEEE Transactions on Solid-State Circuits*, Vol. 11, No. 6, pp. 834–846, Dec. 1976.

[10] P. Leturcq, J. M. Dorkel, A. Napieralski, and E. Lachiver, "A new approach to thermal analysis of power devices," *IEEE Transactions on Electron Devices*, Vol. 34, No. 5, pp. 1147–1156, May 1987.

[11] N. Rinaldi, "Generalized image method with application to the thermal modeling of power devices and circuits," *IEEE Transactions on Electron Devices*, Vol. 49, No. 4, pp. 679–686, April 2002.

[12] Wenhua Dai, P. Roblin, and M. Frei, "Distributed and multiple time-constant electro-thermal modeling and its impact on ACPR in RF predistortion," in *62nd ARFTG Conference*, Boulder, CO, pp. 89–98, 2005.

[13] S. Xu, A. Shibib, Z. Xie, H. Safar, J. Lott, D. Farrel, and M. Mastrapasqua, "High performance RF power LDMOSFET technology for 2.1 GHz power amplifier applications," in *2003 IEEE MTT-S International Microwave Symposium Digest*, 2003.

[14] S. Akhtar, P. Roblin, and J. Strahler, "Electro-thermal and iso-thermal characterizations of power RF silicon LDMOSFETs," *IEEE Transactions on Electron Devices*, Vol. 48, No. 12, pp. 2785–2789, Dec. 2001.

[15] V. M. Dwyer, A. J. Franklin, and D. S. Campbell, "Thermal failure in semiconductor devices," *Solid-State Electronics*, Vol. 33, No. 5, pp. 553–560, May 1990.

[16] P. Hower, C.-Y. Tsai, S. Merchant, T. Efland, S. Pendharkar, R. Steinhoff, and J. Brodsky, "Avalanche-induced thermal instability in LDMOS transistors," in *Proceedings of 2001 International Symposium on Power Semiconductor Devices & ICs*, pp. 153–156, 2001.

[17] Wenhua Dai, "Large Signal Electro-Thermal LDMOSFET Modeling and the Thermal Memory Effects in RF Power Amplifiers," Dissertation, Doctor of Philosophy, Ohio State University, 2004.

[18] A. L. Caviglia and A. A. Iliadis, "Linear dynamic self-heating in SOI MOSFET's," *IEEE Electron Device Letters*, Vol. 14, No. 3, pp. 133–135, Mar. 1993.

[19] F. Cappelluti, F. Bonani, S. Donati Guerrieri, G. Ghione, M. Peroni, A. Cetronio, and R. Graffitti, "A new dynamic, self-consistant electro-thermal model of power HBTs and a novel interpretation of thermal collapse loci in multi-finger devices," in *Proceedings of the IEEE 2001 Custom Integrated Circuits Conference*, pp. 397–400, 2001.

[20] W. R. Curtice, J. A. Plá, D. Bridges, T. Liang, and E. E. Shumate, "A new dynamic electro-thermal nonlinear model for silicon RF LDMOS FETs," in *1999 IEEE MTT-S International Microwave Symposium Digest*, 1999.

[21] O. Jardel, F. De Groote, T. Reveyrand, J.-C. Jacquet, C. Charbonniaud, J.-P. Teyssier, D. Floriot, and R. Quéré, "An electrothermal model for AlGaN/GaN power HEMTs including trapping effects to improve large-signal simulation results on high VSWR," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 55, No. 12-II, pp. 2660–2669, Dec. 2007.

[22] G. J. Riedel, J. W. Pomeroy, K. P. Hilton, J. O. Maclean, D. J. Wallis, M. J. Uren, T. Martin, and M. Kuball, "Nanosecond timescale thermal dynamics of AlGaN/GaN electronic devices," *IEEE Electron Device Letters*, Vol. 29, No. 5, pp. 416–418, April 2008.

[23] J. Faraj, F. De Groote, J.-P. Teyssier, J. Verspecht, R. Quéré, and R. Aubry, "Pulse profiling for AlGaN/GaN HEMTs with large signal characterizations," in *38th European Microwave Conference EuMC 2008*, pp. 757–760, 2008.

[24] S. J. Doo, P. Roblin, G. H. Jessen, R. C. Fitch, J. K. Gillespie, N. A. Moser, A. Crespo, G. Simpson, and J. King, "Effective suppression of *IV* knee walk-out in AlGaN/GaN HEMTs for pulsed-*IV* pulsed-RF with a large signal network analyzer," *IEEE Microwave and Wireless Components Letters*, Vol. 16, No. 12, pp. 681–683, Dec. 2006.

[25] J. Scott, J. G. Rathmell, A. Parker, and M. Sayed, "Pulsed device measurements and applications," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 44, No. 12, pp. 2718–2723, Dec. 1996.

[26] S. J. Doo, P. Roblin, S. Lee, D. Chaillot, and M. Vanden Bossche, "Pulsed-*IV* pulsed-RF measurements using a large signal network analyzer," in *65th ARFTG Conference Digest*, Long Beach, CA, 2005.

[27] S. J. Doo, "New Pulsed-*IV* Pulsed-RF Measurement Techniques for Characterizing Power FETs for Pulsed-RF Power Amplifier Design," Dissertation, Doctor of Philosophy, Ohio State University, 2008.

[28] K. Kunihiro and Y. Ohno, "A large-signal equivalent circuit model for the substrate-induced drain-lag phenomena in HJFET's," *IEEE Transactions on Electron Devices*, Vol. 43, No. 9, pp. 1336–1342, Sept. 1996.

[29] C. K. Yang, P. Roblin, F. D. Groote, J. P. Teyssier, S. Ringel, C. Poblenz, Y. Pei, J. Speck, and U. Mishra, "Pulsed-*IV* pulsed-RF cold-FET parasitic extraction of biased AlGaN/GaN HEMTs using large signal network analyzer," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 58, No. 5-I, pp. 1077–1088, May 2010.

[30] J. P. Ibbetson, P. T. Fini, K. D. Ness, S. P. DenBaars, J. S. Speck, and U. K. Mishra, "Polarization effects, surface states, and the source of electrons in AlGaN/GaN heterostructure field effect transistors," *Applied Physics Letters*, Vol. 77, No. 2, pp. 250–252, July 2000.

[31] R. Vetury, "Polarization induced 2DEG in AlGaN/GaN HEMTs: on the origin, DC, and transient characterization," Dissertation, Doctor of Philosophy, University of California, Santa Barbara, 2000.

[32] A. F. M. Anwar, R. T. Webster, and K. V. Smith, "Bias induced strain in AlGaN/GaN heterojunction field effect transistors and its implications," *Applied Physics Letters*, Vol. 88, No. 5, pp. 2035101–3, May 2006.

[33] C.-K. Yang, P. Roblin, A. Malonis, A. Arehart, S. Ringel, C. Poblenz, Y. Pei, J. Speck, and U. Mishra, "Characterization of traps in AlGaN/GaN HEMTs with a combined large signal network analyzer/deep level optical spectrometer system," in *IEEE International Microwave Symposium*, 2009.

[34] M. Miguez and A. Arnaud, "A study of flicker noise in MOS transistor under switched bias condition," *Journal of Integrated Circuits and Systems*, Vol. 3, No. 2, pp. 63–68, Feb. 2008.

[35] A. S. Roy and C. C. Enz, "Analytical modeling of large-signal cyclo-stationary low-frequency noise with arbitrary periodic input," *IEEE Transactions on Electron Devices*, Vol. 54, No. 9, pp. 2537–2545, Sept. 2007.

[36] I. Suh, P. Roblin, Y. Ko, C.-K. Yang, A. Malonis, A. Arehart, S. Ringel, C. Poblenz, Y. P. Speck, and U. Mishra, "Additive phase noise measurements of AlGaN/GaN HEMTs using a large signal network analyzer and tunable monochromatic light source," in *74th ARFTG Microwave Measurement Symposium*, pp. 1–5, 2009.

[37] J. Breitbarth and J. Koebel, "Additive (residual) phase noise measurement of amplifiers, frequency dividers and frequency multipliers," *Microwave Journal*, Vol. 51, No. 6, pp. 66–82, June 2008.

[38] C. T. Sah, "Theory of low-frequency generation noise in junction-gate field effect transistor," *Proceedings of the IEEE*, Vol. 52, No. 7, pp. 795–814, July 1964.

[39] S. L. Rumyantsev, N. Pala, M. Shur, E. Borovitskaya, A. Dmitriev, M. E. Levinshtein, R. Gaska, M. A. Khan, J. Yang, X. Hu, and G. Simin, "Generation–recombination noise in GaN/AlGaN heterostructure field effect transistors," *IEEE Transactions on Electron Devices*, Vol. 48, No. 3, pp. 530–534, Mar. 2001.

5 Interactive loadline-based design of RF power amplifiers[1]

The design of high-efficiency RF power amplifiers is of critical importance for wireless mobile devices and basestations. Design techniques that rely on circuit simulators are limited by the accuracy of the nonlinear large-signal models and by device package parasitics. Designing amplifiers using load-pull measurements is an alternative approach that bypasses the need for accurate device models. This chapter is concerned with the NVNA-assisted load-pull design of RF power amplifiers.

In this chapter, we will first compare the ideal and measured current and voltage waveforms, dynamic loadlines, and dynamic transfer characteristics for transistors operating in various fundamental amplifier classes (A–F). A multi-harmonic real-time active load-pull (RTALP) technique for the interactive design of the output-matching network will also be introduced and applied to the optimization of the amplifier power efficiency. The aim is to demonstrate how NVNAs can facilitate the design of power amplifiers by providing designers with greater insights into the actual mode of operation of the active devices.

5.1 Review of power amplifiers of various classes (A–F)

Class-A amplifiers provide a tradeoff between linearity and power efficiency. Specifically, they offer an excellent linearity at the price of poor efficiency because the device is on even when no RF signal is applied. The conceptual class-A circuit is shown in Figure 5.1 (a). An RFC is used to provide the supply voltage V_{DD} to the transistor while blocking the RF from the DC supply circuit. The RF signal is directly applied to the gate of the transistor (here an FET). The output RF signal is delivered to the load R_L via a large DC-blocking capacitor of capacitance C.

Shown in the various quadrants of Figure 5.2 are, starting from the top center and going counterclockwise, (a) the ideal I–V characteristics I_D versus V_{DS} for various V_{GS} and the dynamic loadline $i_D(t)$ versus $v_{DS}(t)$, (b) the I–V characteristic I_D versus V_{GS} and the dynamic transfer characteristic $i_D(t)$ versus $v_{GS}(t)$, (c) the $v_{GS}(t)$ waveform, (d) the $v_{DS}(t)$ waveform, and (e) the $i_D(t)$ waveform.

[1] Research collaboration with Xian Cui, Seok Joo Doo, Young Seo Ko, and Wan Rone Liou is gratefully acknowledged.

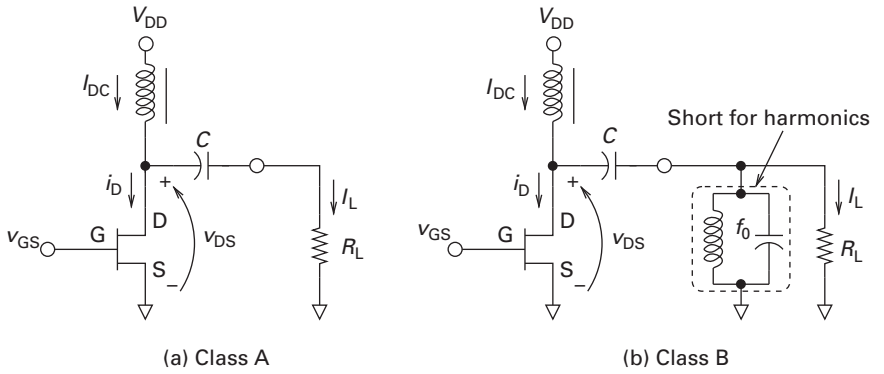


Figure 5.1 Conceptual circuits for class-A (a) and class-B (b) amplifiers.

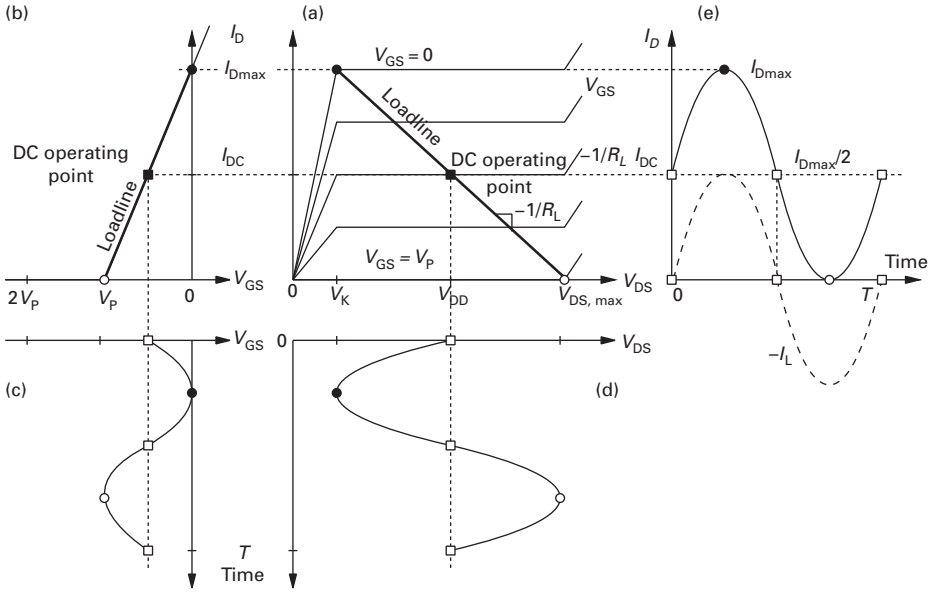


Figure 5.2 Ideal voltage and current waveforms, *I–V* characteristics, dynamic loadline, and transfer characteristic for an FET in class-A operation (see the text).

The voltage at the drain swings between the knee voltage V_K and the maximum drain voltage $V_{DS,max}$, which is up to twice the DC voltage supply V_{DD}. Note that a knee voltage V_K is accounted for in the idealized *I–V* characteristics. The optimal load R_L for maximum output power for this ideal transistor is then given by

$$R_L = \frac{V_{Dmax} - V_K}{I_{Dmax}},$$

with I_{Dmax} and V_{Dmax} the maximum current and voltage with which the device and circuits can be safely operated. This simple model obviously neglects the capacitances exhibited by the devices, and we will revisit this topic in the next section.

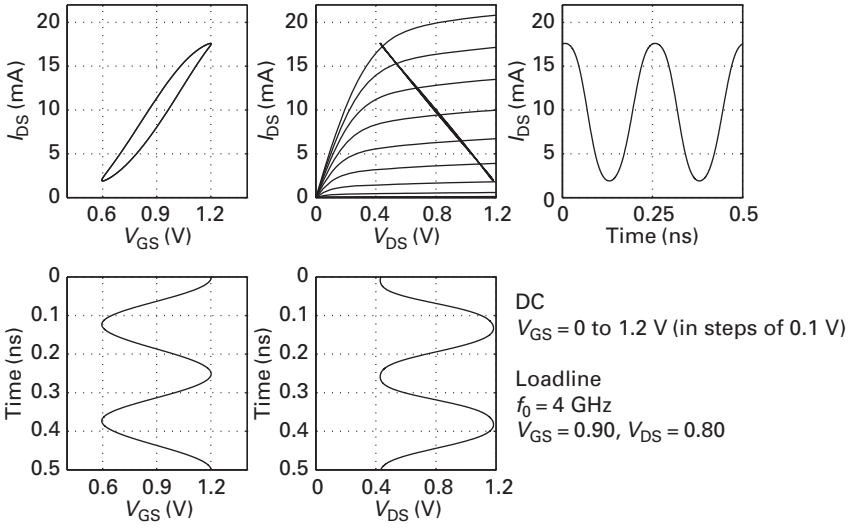


Figure 5.3 Measured voltage and current waveforms, *I–V* characteristics, dynamic loadline, and transfer characteristic for an FET (MOSFET) in class-A operation.

Measurement results acquired with an LSNA and an *I–V* tracer on an n-channel MOSFET operating in class A for a signal at 4 GHz and a load resistance of about 50 Ω are shown in Figure 5.3. The various quadrants again show the *I–V* characteristics, dynamic loadlines, dynamic transfer characteristic, and current and voltage waveforms. The device performance is similar to that of the ideal device except for the $i_D(t)$–$v_{GS}(t)$ dynamic transfer characteristic, which reveals a lag between the gate voltage and the drain current. This lag results, among other things, from the drain-to-source capacitances and associated *R–C* charging time constants. Note that the drain current and drain voltage are in phase since a broadband resistance R_L for the fundamental and harmonics was used as a load for this class-A measurement.

The power efficiency in the case of an ideal transistor with $V_K = 0$ is graphically analyzed in Figure 5.4. For class A the RF output power is

$$P_{RF} = \frac{I_D^2(\omega_0)R_L}{2} = \frac{R_L}{2}\left(\frac{I_{Dmax}}{2}\right)^2 = \frac{I_{Dmax}^2 R_L}{8},$$

where $I_D(\omega_0)$ is the fundamental component of the drain current. This corresponds to the top triangle shown in Figure 5.4. The DC power dissipation is itself given by the product of the average (DC) current and the average (DC) drain voltage:

$$P_{DC}(\text{Class A}) = I_{DC} \times V_{DD} = \frac{I_{Dmax}}{2} \times \frac{R_L I_{Dmax}}{2} = \frac{I_{Dmax}^2 R_L}{4}.$$

This leads to a drain efficiency for an ideal transistor with $V_K = 0$ in class A of

$$\eta(\text{Class A}) = \frac{P_{RF}}{P_{DC}} = \frac{1}{2} = 50\%.$$

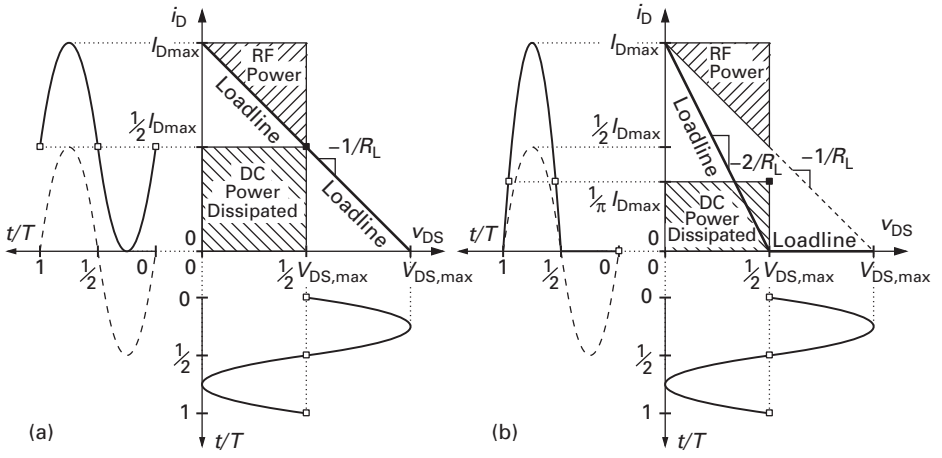


Figure 5.4 Graphical comparison of power dissipation in class-A (a) and class-B (b) amplifiers

Under power-backoff operation the DC power is the same for class A but the RF power is reduced because the fundamental RF current $I_D(\omega_0)$ is now smaller than $I_{Dmax}/2$. The efficiency under power backoff scales therefore as

$$\eta(\text{Class A}) = 50\% \times \left(\frac{I_D(\omega_0)}{I_{Dmax}/2}\right)^2.$$

The knee voltage will also reduce the efficiency of the PA since it is equivalent to an effective power backoff. The RF voltage swing is indeed limited to $V_{DS,max} - V_K$ and thus the current amplitude by $I_D(\omega_0) = (V_{DS,max} - V_K)/(2R_L)$ if we assume that the same load R_L is applicable. In real devices the knee voltage increases with increasing gate voltages, leading to further decrease in the efficiency. An improved efficiency is therefore usually observed at lower RF frequencies when operating with a lower maximum drain current I_{Dmax}.

The class-B amplifier provides an enhanced efficiency compared with class A but is in practice substantially more nonlinear. In class B, the transistor is biased so that it is conducting during only half of the RF cycle (see Figure 5.5). As shown in Figure 5.1(b) the transistor is loaded with a resonator, which ideally shorts all the harmonics so that the drain voltage is purely sinusoidal despite the rectified drain current. It results that the dynamic loadline is piecewise linear. The resistance at the fundamental is, however, still given by $R_L = (V_{Dmax} - V_K)/I_{Dmax}$. The advantage of the class-B amplifier is that the drain current is zero when no RF power is applied such that the amplifier does not ideally dissipate power. When a sufficiently large RF power is applied at the gate for rail-to-rail operation, the DC drain current rises to I_{Dmax}/π. The DC power dissipated in class B for an ideal transistor with $V_K = 0$ is therefore given by

$$P_{DC}(\text{Class B}) = I_{DC} \times V_{DD} = \frac{I_{Dmax}}{\pi} \times \frac{I_{Dmax} R_L}{2} = \frac{I_{Dmax}^2 R_L}{2\pi}.$$

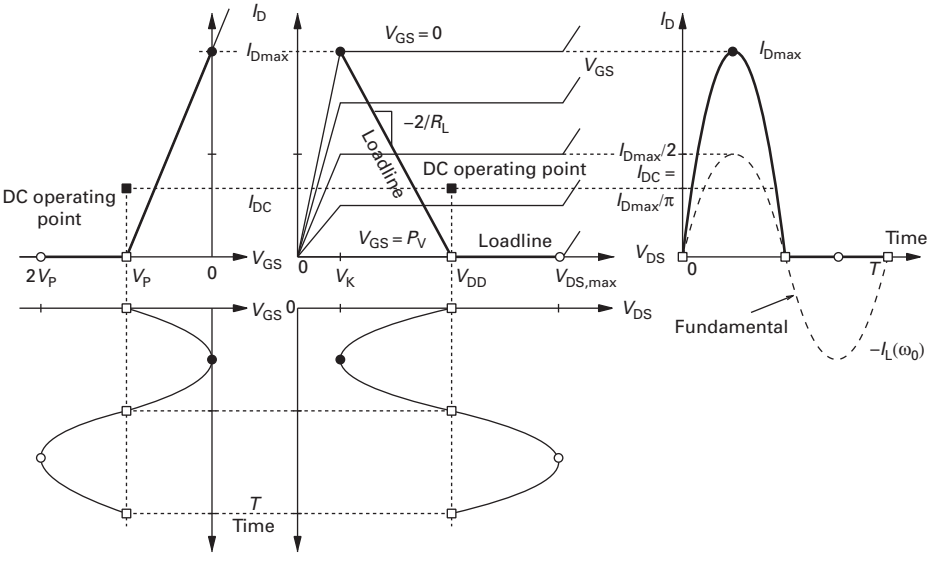


Figure 5.5 Ideal voltage and current waveforms, *I–V* characteristics, dynamic loadline, and dynamic transfer characteristic, for an FET in class-B operation.

The DC power provided is indicated in Figure 5.4(b) by a cross-hatched square. The output RF power provided by the transistor is still given by $P_{RF}(\text{Class B}) = I_{Dmax}^2 R_L/8$ as in class A, as indicated by the top triangle in Figure 5.4(b). It results that the drain efficiency for an ideal transistor is given by

$$\eta(\text{Class B}) = \frac{P_{RF}}{P_{DC}} = \frac{2\pi}{8} = \frac{\pi}{4} = 78.5\%.$$

This corresponds to an important efficiency improvement over class A. Under power-backoff operation both the DC power and the RF power are reduced since the fundamental RF current $I_D(\omega_0)$ is now smaller than $I_{Dmax}/2$. The efficiency under power backoff scales therefore as

$$\eta(\text{Class B}) = \frac{\pi}{4}\frac{I_{Dmax}}{2I_D(\omega_0)}\left(\frac{I_D(\omega_0)}{I_{Dmax}/2}\right)^2 = 78.5\% \times \frac{I_D(\omega_0)}{I_{Dmax}/2}.$$

Thus the efficiency in class B derates under power backoff at a slower rate than for class A. Measurements performed with an LSNA and *I–V* tracer for an n-channel MOSFET operating in class B are shown in Figure 5.6 for a signal at 4 GHz and a load resistance of about 50 Ω. The second harmonic was shorted by injecting an 8-GHz signal at the output. The various quadrants again show the *I–V* characteristics, dynamic loadline, dynamic transfer characteristic, and current and voltage waveforms. The device performance is similar to that of the ideal device but again a small lag of $i_D(t)$–$v_{GS}(t)$ is observed in the dynamic loadline due to the drain-to-source capacitance. Also, due to the finite number of harmonics measured, ringing in the drain current is observed when the device turns off.

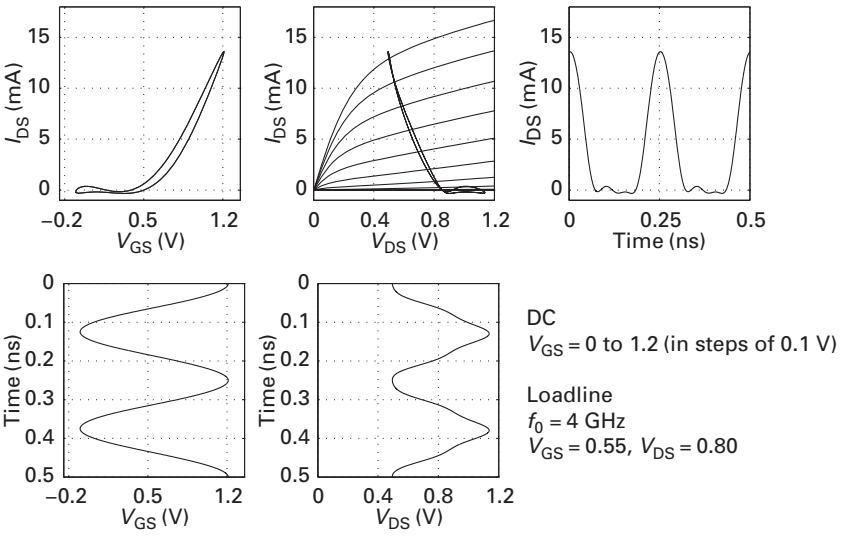


Figure 5.6 Measured voltage and current waveforms, *I–V* characteristics and dynamic loadline, and transfer characteristics for an FET (MOSFET) in class-B operation.

Class-A and class-B operation correspond to drain-current duty cycles of 100% and 50%, respectively. Note that the drain-current duty cycle is defined as the fraction of the RF period for which the drain current is non-zero (transistor on).

The ideal class-B amplifier is linear. In practice, however, for gate voltages near the threshold (pinchoff) voltage the transconductance reduces (subthreshold region) and the device operates nonlinearly. As a consequence, amplifiers are operated in class AB rather than class B, where the duty cycle is between 50% and 100% but usually close to 50%. Class C is associated with duty cycles smaller than 50%.

The variation of the output power relative to class-A operation and the efficiency are both plotted as a function of the duty cycle in Figure 5.7 for an ideal transistor with $V_K = 0$. The reader is referred to Ref. [1] for the derivation. An efficiency of 100% is theoretically achievable in class C for vanishing duty cycles, but at the price of reduced output power for a given $V_{DS,max}$ and I_{Dmax}. The DC current I_{DC} and fundamental components of the drain current $I_D(\omega_0)$ are also plotted as a function of the duty cycle in Figure 5.8 [1].

Class-F amplifiers provide further improvements over class B in terms of efficiency. The general circuit is shown in Figure 5.9(a). The circuit makes use of the load network to shape the voltage waveform so as to increase the power efficiency. In its ultimate conceptual realization (class D), the drain-to-source voltage waveform is shaped by the load circuit to generate a square wave in response to the half sine wave of the drain current as is shown in Figure 5.10. The power efficiency then reaches 100% for the ideal transistor with $V_K = 0$ since the drain-to-source voltage is zero when the drain current rises. For class-F operation the load network must provide an open circuit for the odd harmonics and a short circuit for the even harmonics. The load resistance required at the fundamental frequency is itself given by

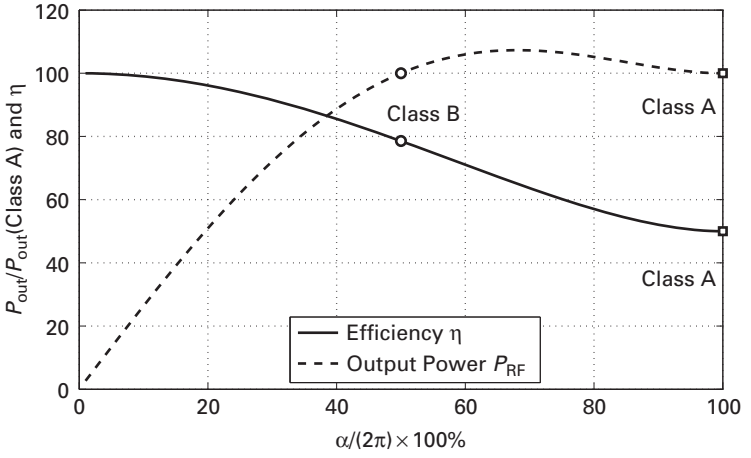


Figure 5.7 Efficiency and normalized output power relative to class A (100% conduction angle) versus the conduction angle.

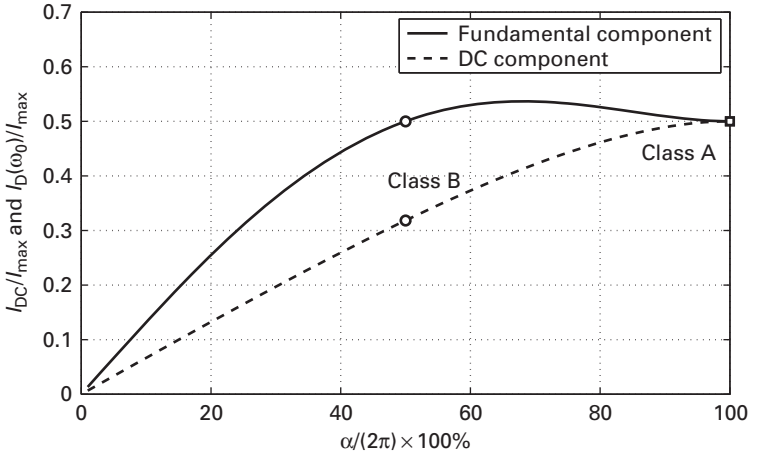


Figure 5.8 The variation of the DC component and the fundamental component as a function of the conduction angle.

$$R_{\mathrm{L}}(\text{Class F}) = \frac{V_{\mathrm{D}}(\omega_0)}{I_{\mathrm{D}}(\omega_0)} = \frac{4}{\pi} V_{\mathrm{DD}} \times \frac{2}{I_{\mathrm{Dmax}}} = \frac{4}{\pi} R_{\mathrm{L}}(\text{Class A}).$$

NVNA measurements for real transistors operating in class F will be presented in Sections 5.3–5.5.

Class E amplifiers also theoretically achieve 100% efficiency by using a transistor as a switch, to charge and discharge (when the switch is off) a capacitance C_{P} placed in parallel between drain and source as shown in Figure 5.9(b). A series resonator is used to maintain the current purely sinusoidal by providing an open circuit for the harmonics.

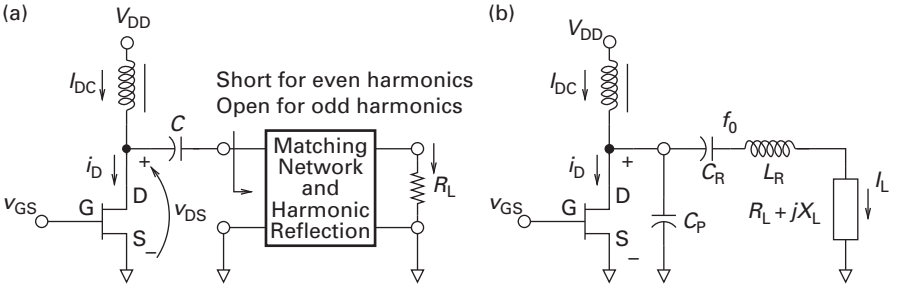


Figure 5.9 Conceptual circuits for class-F (a) and class-E (b) amplifiers.

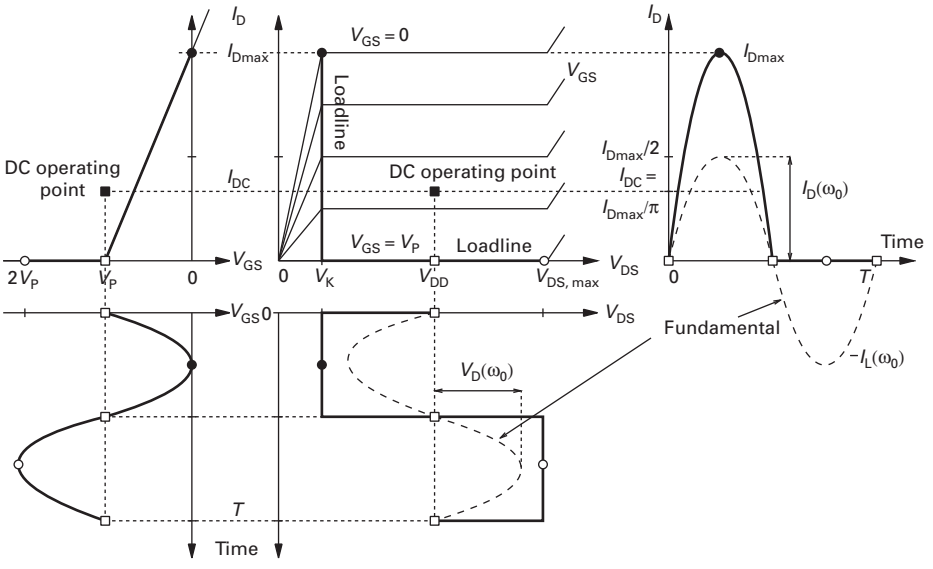


Figure 5.10 Ideal voltage and current waveforms, *I–V* characteristics, dynamic loadline, and dynamic transfer characteristic for an FET in class-F operation.

The resulting current and voltage waveforms are shown in Figure 5.11. It is critical that the capacitor be fully discharged ($V_{DS} = 0$) when the switch turns on in order to achieve maximum efficiency. For a design with a duty cycle of 50% this leads to the following design parameters [2] [3]:

$$R_L = 0.577\frac{(V_{DD} - V_K)^2}{P_{RF}},$$
$$C_P = \frac{0.1837}{\omega_0 R_L},$$
$$X_L = 1.152R_L,$$

beside the resonant condition $\omega_0 = 1/(L_R C_R)$. Experimental measurements with an NVNA are presented in Section 5.7.

Table 5.1. Impedance termination of the fundamental and harmonics in various classes. The definition $R_L(A) = V_{Dmax}/I_{Dmax}$ is used for reference.

	Impedance		
Classes	Fundamental	Odd harmonics	Even harmonics
A	$R_L(A)$	None	None
AB–B–C	$R_L(A)$	Short	Short
F (D)	$(4/\pi)R_L(A)$	Open	Short
Inverted F (D)	$(\pi/4)R_L(A)$	Short	Open
E	$0.464R_L(A)$	Open	Open

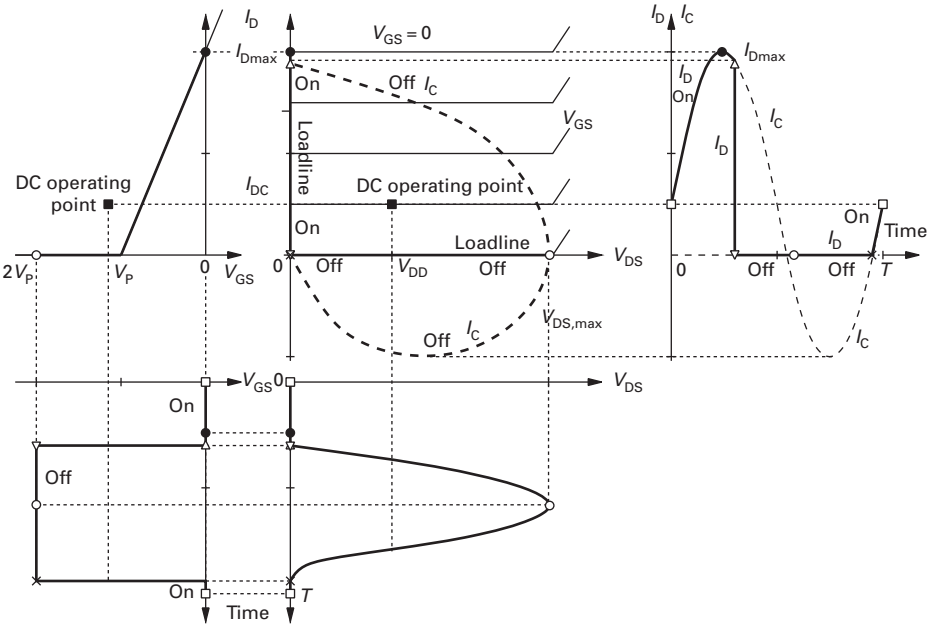


Figure 5.11 Ideal voltage and current waveforms, *I–V* characteristics, dynamic loadline, and dynamic transfer characteristic for an FET in class-E operation.

The fundamental and harmonic impedance terminations needed for the various classes discussed are summarized in Table 5.1 for the case of a perfect transistor. Clearly the impedance termination of the harmonics has an important impact on the device operation.

To further unify the analysis of amplifiers of various classes Ref. [4] introduced a set of four factors that establish the proportionality between the DC current and voltage and (1) the amplitude and (2) the maximum value of the fundamental voltage and current waveforms:

Table 5.2. Values of the γ and δ parameters for various classes. Also shown are the efficiency and power-output capability figures of merit. For class E the two sets of values E(1) and E(2) are for switching with a duty cycle of 50% without and with second-harmonic peaking [3] [4].

Class	γ_V	γ_I	δ_V	δ_I	η	$\eta/(\delta_V \delta_I)$
A	1	1	2	2	1/2	0.125
B	1	$\pi/2$	2	π	$\pi/4$	0.125
F	$4/\pi$	$\pi/2$	2	π	1	0.159
i-F	$\pi/2$	$4/\pi$	π	2	1	0.159
E(1)	1.638	1.847	3.56	2.84	1	0.098
E(2)	1.414	1.414	2.912	2.912	1	0.118

$$V_{\mathrm{D}}(\omega_0) = \gamma_V V_{\mathrm{DD}},$$
$$V_{\mathrm{Dmax}} = \delta_V V_{\mathrm{DD}},$$
$$I_{\mathrm{D}}(\omega_0) = \gamma_I I_{\mathrm{DC}},$$
$$I_{\mathrm{Dmax}} = \delta_I I_{\mathrm{DC}}.$$

These factors are given in Table 5.2 for the various classes. For a general complex load $Z_{\mathrm{L}} = R_{\mathrm{L}} + jX_{\mathrm{L}}$ the output RF power in the real load then reduces to

$$P_{\mathrm{RF}} = \frac{1}{2}\,\mathrm{Re}\left[\frac{|V_{\mathrm{D}}(\omega)|^2}{Z_{\mathrm{L}}}\right] = \frac{\gamma_V^2 V_{\mathrm{DD}}^2}{2\rho^2 R_{\mathrm{L}}} \qquad \text{with} \quad \rho = \frac{|Z_{\mathrm{L}}|}{R_{\mathrm{L}}},$$

and the drain efficiency is

$$\eta = \frac{P_{\mathrm{RF}}}{P_{\mathrm{DC}}} = \frac{\gamma_V \gamma_I}{2\rho}.$$

When the drain current and the drain-to-source voltage are 180° out of phase as in classes A, B, C, and F, the output RF power in the resistive load $Z_{\mathrm{L}} = R_{\mathrm{L}}$ reduces simply to

$$P_{\mathrm{RF}} = \frac{1}{2}\gamma_V \gamma_I I_{\mathrm{DC}} V_{\mathrm{DD}} = \frac{1}{2}\gamma_V \gamma_I P_{\mathrm{DC}}.$$

For classes A, B, C, and F the drain efficiency is then simply

$$\eta = \frac{\gamma_V \gamma_I}{2}.$$

Another figure of merit is the power-output capability introduced in Ref. [4] as

$$\frac{P_{\mathrm{RF}}}{V_{\mathrm{Dmax}} I_{\mathrm{Dmax}}} = \frac{P_{\mathrm{RF}}}{P_{\mathrm{DC}}} \times \frac{P_{\mathrm{DC}}}{V_{\mathrm{Dmax}} I_{\mathrm{Dmax}}} = \frac{\eta}{\delta_V \delta_I}.$$

As shown in Table 5.2, the class-F amplifier slightly outperforms class-A–B and class-E amplifiers, by 27% and 35%, respectively, using this technology-oriented figure of merit.

Amplifier classes have been presented in this section for the case of an ideal transistor with constant transconductance and no device capacitances, no parasitics, no knee voltage, and no on (drain) resistance. The impedance terminations derived for the various classes of operation in the ideal-transistor case will therefore need to be modified when applied to real transistors. Given the complexity of real transistors, a common design approach relies on the systematic sweep of the fundamental and harmonic impedances to establish the optimal performance. However, as we shall see, for the real transistor just as for the ideal transistor, lossless impedance terminations for the harmonics do yield the best performance for the power efficiency in each of the classes considered.

In addition it was noted that maximum efficiency in classes E and F was obtained for the perfect transistor by achieving non-overlapping drain current and drain-to-source voltage waveforms such that we have

$$v_{DS}(t)i_D(t) = 0 \qquad \text{for all time } t \text{ in the RF cycle.} \tag{5.1}$$

In real devices this target can be difficult to approach since more complex waveforms may arise due to the device parasitics and nonlinear capacitances. In such a case, we shall see that a more effective goal to maximize the efficiency for a given P_{DC} is to minimize the average power dissipated by the transistor over the RF cycle of duration T:

$$P_{diss} = \frac{1}{T}\int_0^T v_{DS}(t)i_D(t)dt. \tag{5.2}$$

This is exemplified by the ideal class-E amplifier, where the parallel capacitance C_P is storing energy during a fraction of the off-cycle and releasing energy during another fraction of the off-cycle. It is therefore beneficial if the instantaneous power $v_{DS}(t)i_D(t)$ can be negative (generation) for a substantial fraction of the RF cycle to compensate for the region of positive (dissipation) instantaneous power. Maximum efficiency will therefore be obtained when the current and voltage waveforms are orthogonal to each other as specified by Equation (5.2) rather than enforcing the zero cross product of Equation (5.1).

To be rigorous, we should also include the power dissipated in the gate, and the total dissipated power is generalized to

$$P_{diss} = \frac{1}{T}\int_0^T [v_{GS}(t)i_G(t) + v_{DS}(t)i_D(t)]dt.$$

Similar relations can be written for a bipolar transistor.

Note also that, in real transistors, the impedance termination provided by the input matching network for the harmonics or the waveform (sinusoidal or non-sinusoidal) used for the input RF excitation can also impact the performance of the amplifier.

5.2 Output termination with load-pull measurements

Load-pull is a measurement and a simulation technique that consists of varying the load impedance seen by a transistor at the fundamental or harmonics while measuring

the performance of the transistor. Load-pull can be used to measure a transistor under realistic operating conditions, and it helps the designer determine the optimum load impedance Z_L which delivers the maximum power or maximum efficiency. Furthermore, a load-pull system used jointly with an NVNA enables the designer to visualize the voltage and current waveforms at the DUT planes and plot the dynamic loadline and transfer characteristic of the transistor.

Traditionally load-pull and source-pull have been realized using passive tuners. In the simplest case these tuners are manually operated, whereas in the most sophisticated case they are computer-controlled. Multi-harmonic tuners allowing one to set simultaneously the impedances at the fundamental and second and third harmonic frequencies are now available.

With the advent of NVNAs, new types of load-pulls have also emerged. Two different load-pull techniques will be considered in this section: active load-pull and real-time active load-pull measurements.

5.2.1 Active load-pull measurements

The systematic design of RF power amplifiers has been limited because of a general lack of impedance data under actual operating conditions. To overcome this limitation encountered in passive load-pull measurements, an active load-pull technique was introduced by Y. Takayama in [5]. Several active load-pull measurement schemes have since been developed. Active load-pull systems relying on the closed-loop architecture [6] [7] improve on the reflection-coefficient range, but are subject to potential stability problems. The open-loop active load-pull system [8] overcomes the stability issue. Since active load-pull measurements use an active loop or rely on the injection of RF power from additional external signal sources, active load-pull systems can easily overcome the loss problem in passive tuners. Indeed, they permit one to generate a reflection coefficient with an amplitude equal to or even larger than unity.

Figure 5.12 shows a harmonic active load-pull system (ALP) using three RF signal generators. The harmonic reflection coefficients are defined as

$$\Gamma_L(n\omega_0) = \frac{a_2(n\omega_0)}{b_2(n\omega_0)}, \tag{5.3}$$

where n = 1, 2, and 3. Each phase shifter and amplifier changes the phase and amplitude of the corresponding harmonic signal $a_2(n\omega_0)$, so that the targeted reflection-coefficients $\Gamma_L(n\omega_0)$ are produced. Actually, with modern synthesized RF signal sources, the phase of the RF signal generated can be easily tuned. The various RF sources need then to be locked to the same time-base reference (typically 10 or 100 MHz) so that the phase drift is negligible. In such a case no external phase shifter is needed. Isolators are, however, required in practice in ALP to protect the various RF sources and their amplifiers from the power generated by the DUT.

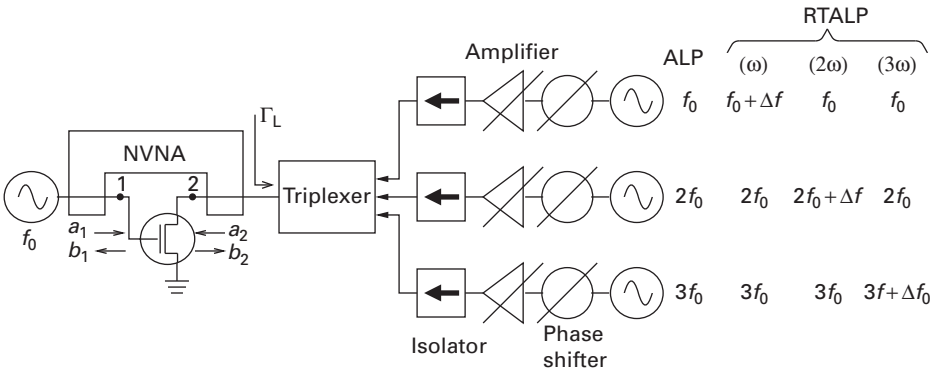


Figure 5.12 The concept of an active harmonic load-pull system (ALP) and real-time active load-pull (RTALP).

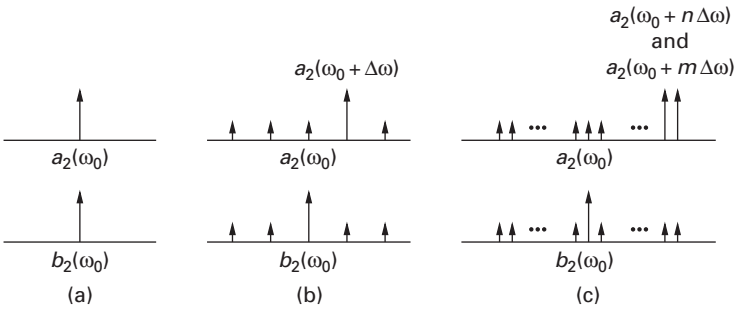


Figure 5.13 Comparison of several active load-pull schemes at ω_0: (a) traditional active load-pull, (b) real-time active load-pull, and (c) ultra-fast real-time active load-pull. In (c), $m = n + 1$.

5.2.2 Real-time active load-pull measurements

In general, load-pull (passive or active) measurements are very time-consuming, since a lot of information is needed for the determination of the optimum load termination.

A real-time active load-pull (RTALP) technique using the LSNA that has the potential for drastically reducing the PA design cycle was recently introduced [9]. RTALP, which was first demonstrated for the fundamental frequency, can also be extended to second- and third-harmonic load-pull measurements for the interactive design of class-F nonlinear power amplifiers [10], leading to the testbed implementation shown in Figure 5.12.

Figure 5.13 compares three different active load-pull methods in the frequency domain. In a departure from the traditional active load-pull measurements shown in Figure 5.13(a), in the real-time active load-pull methods shown in Figure 5.13(b) an incident signal $a_2(\omega_0 + \Delta\omega)$ is injected at port 2 (drain) with a slightly different frequency $\omega_0 + \Delta\omega$ from the frequency ω_0 injected at port 1 (gate). The resulting frequency offset $\Delta\omega$ is equivalent to a phase sweeping $\phi = \Delta\omega\, t$ of the reflected wave $a_2(\omega_0, t)$ at the frequency ω_0. As the phase of $a_2(\omega_0, t)$ is effectively swept, the reflection coefficient

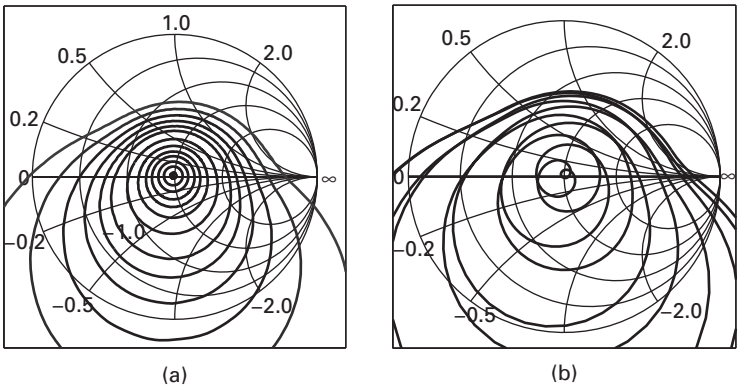


Figure 5.14 Comparison of $\Gamma_L(\omega_0, t)$ obtained from (a) real-time active load-pull and (b) ultra-fast real-time active load-pull.

$\Gamma_L(\omega_0, t) = a_2(\omega_0, t)/b_2(\omega_0, t)$ takes a continuum of values along a closed-loop trajectory in the extended Smith chart. Effectively in a single LSNA measurement a large set of load reflection values are simultaneously acquired this way.

The time-dependent reflection coefficient $\Gamma_L(n\omega_0, t)$ for the nth harmonic can be extracted from the following equation based on the RTALP measurements:

$$\Gamma_L(n\omega_0, t) = \sum_{m=-M}^{M} a_2(n\omega_0 + m\,\Delta\omega)e^{jm\Delta\omega t} \Big/ \sum_{m=-M}^{M} b_2(n\omega_0 + m\,\Delta\omega)e^{jm\Delta\omega t},$$

where M indicates the number of SSB modulation tones acquired by the LSNA. An example of $\Gamma_L(n\omega_0, t)$ from the RTALP method is shown in Figure 5.14(a). In order to cover the entire Γ_L plane on the Smith chart, the RTALP method requires the variation of the amplitude of the injected signal $a_2(\omega_0 + \Delta\omega)$.

The time-varying output power can be obtained from multitones by using

$$P_{out}(\omega_0, t) = -\frac{1}{2}\sum_{p=-M}^{M}\sum_{q=-M}^{M}\left(v_2^p \times i_2^q\right)e^{j(p-q)\Delta\omega t}, \tag{5.4}$$

where $v_2^p = v_2(\omega_0 + p\,\Delta\omega)$ and $i_2^q = i_2^*(\omega_0 + q\,\Delta\omega)$. The time-varying PAE is then obtained using

$$\mathrm{PAE}(\omega_0, t) = \frac{P_{out}(\omega_0, t) - P_{in}(\omega_0, t)}{P_{DC}(t)}. \tag{5.5}$$

Having presented the multitone definition of the reflection coefficient, output power, and efficiency, further insights can be gained by limiting the analysis to the two dominant tones, namely $n\omega_0$ and $n\omega_0+\Delta\omega$. The incident and reflected waves at port 2 can be written as

$$A_{2,n}(t) = a_2(n\omega_0)\exp(jn\omega_0 t) + a_2(n\omega_0+\Delta\omega)\exp\left[j(n\omega_0+\Delta\omega)t\right],$$
$$B_{2,n}(t) = b_2(n\omega_0)\exp(jn\omega_0 t) + b_2(n\omega_0+\Delta\omega)\exp\left[j(n\omega_0+\Delta\omega)t\right].$$

The instantaneous reflection coefficient $\Gamma_L(n\omega_0, t)$ can then be verified to be rewritten after some algebra:

$$\Gamma_L(n\omega_0, t) = \frac{A_{2,n}(t)}{B_{2,n}(t)} = \Gamma_{L0}(n\omega_0) + \frac{1 - \Gamma_{L0}(n\omega_0)\Gamma_{out}(n\omega_0+\Delta\omega)}{\Gamma_{out}(n\omega_0 + \Delta\omega)} \times \frac{1}{1 + R_b(n\omega_0, \Delta\omega)\exp(-j\Delta\omega t)}, \tag{5.6}$$

where the following time-independent terms were defined:

$$\Gamma_{L0}(n\omega_0) = \frac{a_2(n\omega_0)}{b_2(n\omega_0)},$$
$$\Gamma_{out}(n\omega_0+\Delta\omega) = \frac{b_2(n\omega_0+\Delta\omega)}{a_2(n\omega_0+\Delta\omega)},$$
$$R_b(n\omega_0, \Delta\omega) = \frac{b_2(n\omega_0)}{b_2(n\omega_0+\Delta\omega)} = \Gamma_{out}^{-1} \times \frac{b_2(n\omega_0)}{a_2(n\omega_0+\Delta\omega)}.$$

$\Gamma_{L0}(n\omega_0)$, is the reflection coefficient of the load at $n\omega_0$, which can be changed using a passive tuner. $\Gamma_{out}(n\omega_0+\Delta\omega)$ is the output reflection of the device under large-signal operation. Some more insights into the ratio $R_b(n\omega_0, \Delta\omega)$ can be obtained if we first consider small-signal excitations. The device being linear for small-signal excitations, the superposition principle holds and the reflected waves $b_2(\omega_0)$ and $b_2(\omega_0+\Delta\omega)$ at port 2 are independent. Similarly, Γ_{out} is also independent of $b_2(\omega_0)$ and the ratio R_b can then be controlled by adjusting the amplitude of $a_2(\omega_0+\Delta\omega)$ at port 2 for a given amplitude of the incident wave $b_2(\omega_0)$ at port 2. In the limit of vanishing $a_2(\omega_0+\Delta\omega)$ the ratio R_b becomes infinite, and the instantaneous reflection coefficient $\Gamma_L(n\omega_0, t)$ reduces to the load reflection coefficient $\Gamma_{L0}(n\omega_0)$. For large-signal excitations $b_2(\omega_0)$ is no longer independent of the incident power $b_2(\omega_0+\Delta\omega)$ at port 2. For example, consider an excitation applied both at port 1 and at port 2 at the fundamental frequencies ω_0 and $\omega_0+\Delta\omega$ with no harmonic injected at port 1 and port 2 (ports perfectly matched). In such a case, using the techniques introduced in Chapter 6, the transmitted wave $b_2(\omega_0 + p\,\Delta\omega)$ at port 2 (with p an integer) is of the form

$$\begin{aligned}
&\cdots \\
b_2(\omega_0 - \Delta\omega) &= V_{-3}(\mathbf{x}_2) \cdot a_1^2(\omega_0)a_2^*(\omega_0 + \Delta\omega), \\
b_2(\omega_0) &= V_{21}(\mathbf{x}_2) \cdot a_1(\omega_0), \\
b_2(\omega_0 + \Delta\omega) &= V_{22}(\mathbf{x}_2) \cdot a_2(\omega_0 + \Delta\omega), \\
b_2(\omega_0 + 2\,\Delta\omega) &= V_3(\mathbf{x}_2) \cdot a_1^*(\omega_0)a_2^2(\omega_0 + \Delta\omega), \\
&\cdots
\end{aligned}$$

where the Volterra functions $V_{-3}(\mathbf{x}_2)$, $V_{21}(\mathbf{x}_2)$, $V_{22}(\mathbf{x}_2)$, and $V_3(\mathbf{x}_2)$ are analytic functions of the multivariable vector $\mathbf{x}_2$ defined (in the absence of input harmonics) as

$$\mathbf{x}_2 = \Big[\omega_0,\ \omega_0 + \Delta\omega, |a_1(\omega_0)|^2, |a_2(\omega_0+\Delta\omega)|^2\Big].$$

Note that the Volterra functions V_{21} and V_{22} reduce to the scattering parameters $S_{21}(\omega_0)$ and $S_{22}(\omega_0 + \Delta\omega)$, respectively, in the limit of low input power at port 1 and port 2. These results indicate that $b_2(\omega_0)$ and $b_2(\omega_0+\Delta\omega)$ are no longer independent, since they are, among other things, both affected by the input power levels $|a_1(\omega_0)|^2$ and $|a_2(\omega_0+\Delta\omega)|^2$. Nevertheless, sufficient degrees of freedom are still available in practice from the transistor, so the ratio R_b can be varied by adjusting the amplitude of $a_2(\omega_0+\Delta\omega)$ at port 2 for a given amplitude of $a_1(\omega_0)$.

The trajectory of $\Gamma_{\mathrm{L}}(n\omega_0, t)$ defined by the two-tone Equation (5.6) inscribes a circle with center $\Gamma_{\mathrm{ALP},n}$ and radius $R_{\mathrm{ALP},n}$ given by

$$\left|\Gamma_{\mathrm{L}}(n\omega_0, t) - \Gamma_{\mathrm{ALP},n}\right| = R_{\mathrm{ALP},n}$$

with

$$\begin{aligned} R_{\mathrm{ALP},n} &= |S_n| \times \left(\Gamma_b^2 - \Gamma_b\right)^{1/2}, \\ \Gamma_{\mathrm{ALP},n} &= \Gamma_{\mathrm{L0}}(n\omega_0) + S_n \times \Gamma_b, \\ \Gamma_b &= \frac{1}{1 - |R_b|^2}, \\ S_n &= \frac{1 - \Gamma_{\mathrm{L0}}(n\omega_0)\Gamma_{\mathrm{out}}(n\omega_0+\Delta\omega)}{\Gamma_{\mathrm{out}}(n\omega_0+\Delta\omega)}. \end{aligned}$$

Examples of such circle trajectories obtained from the two-tone theory using measured data acquired with a real-time active load-pull system are shown in Figure 5.15.

The circle trajectories hold both for small- and for large-signal excitations in the limit where only two tones are considered to reconstruct the incident $A_{2,n}(t)$ and reflected $B_{2,n}(t)$ waves at port 2. However, under large-signal excitations, intermodulation products are also generated by the nonlinear device, which introduces further modulation in the envelope of the reflected waves at port 2. When accounting for these additional tones in the calculation of the incident $A_{2,n}(t)$ and reflected $B_{2,n}(t)$ waves at port 2, the closed-loop trajectory inscribed by $\Gamma_{\mathrm{L}}(n\omega_0, t)$ will depart from a circle. Indeed, the multitone theory (dashed line) is seen to depart from the two-tone theory (plain line) for high input power levels injected at the output (port 2). Nevertheless, valuable insights into the principle of RTALP are gained from the two-tone theory.

Figure 5.13 (c) introduces a more advanced RTALP method that uses two-tone injection at port 2 instead of a single-tone excitation [11]. As illustrated in Figure 5.14(b), the effect of the two-tone excitation is to modulate the envelope of the applied active signal to effectively sweep the power in order to characterize the entire Γ_{L} plane in a single measurement. The total required time for this single measurement is approximately 10 ms as set by the inverse of the 95-Hz resolution bandwidth used for the LSNA.

It is important to note that the accuracy of the load optimization in the design of power amplifiers could be affected by memory effects in the transistor due to the tone

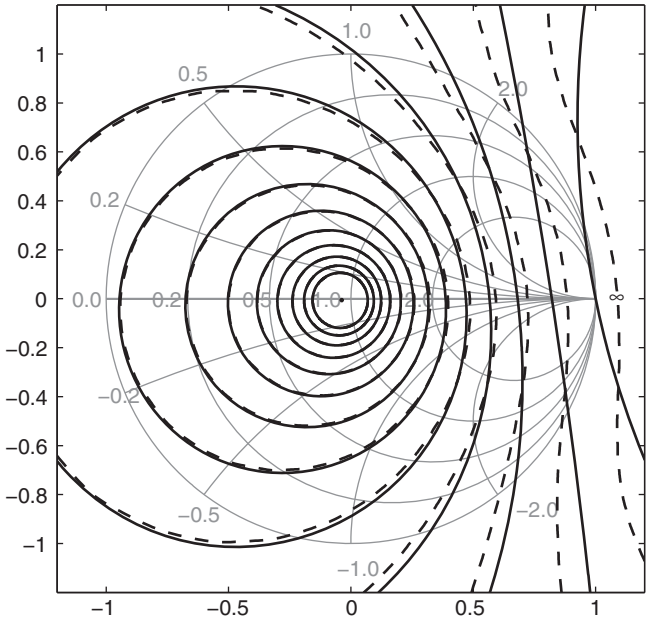


Figure 5.15 Loci of $\Gamma_L(\omega_0)$ obtained from RTALP measurements for various injected power levels at port 2 and displayed using the leading two tones (plain lines) or using all the tones (dashed lines). Excellent agreement is obtained for low injected power levels at port 2.

spacing $\Delta\omega$ used for conducting the real-time active load-pull measurements. However, the accuracy of the optimal loads extracted using RTALP measurement results can be verified by using the constant-phase active-load method. In addition, the modulation with $\Delta\omega$ normally generates low-frequency intermodulation products, and this results in the beating of the DC bias. Synchronized baseband and RF measurements are therefore required. Compensation for the triggering delay between the baseband and RF data acquisition is thus needed in RTALP measurements for accurate PAE measurements. For this purpose a reference time calibration can be first performed as explained in Chapter 2.

5.3 Class-F design with RTALP

In this section we shall consider the example of the design of a class-F device operating at 2 GHz using RTALP [10]. The device is a GaN HEMT biased in class B–F with $V_{DS} = 4.25$ V and $V_{GS} = -2.77$ V. First RTALP measurements are conducted for the fundamental frequency for 14 power levels injected at the output while keeping port 2 matched for the second and third harmonics. The input power level at the fundamental frequency is maintained constant throughout the measurements. It is selected to insure a gate voltage swing of 3.55 V peak to peak. The modulation frequency (tone spacing) selected is 200 kHz and the resolution frequency is 95 Hz, corresponding to about 10 ms of acquisition time. Four harmonics are acquired. The results are shown in

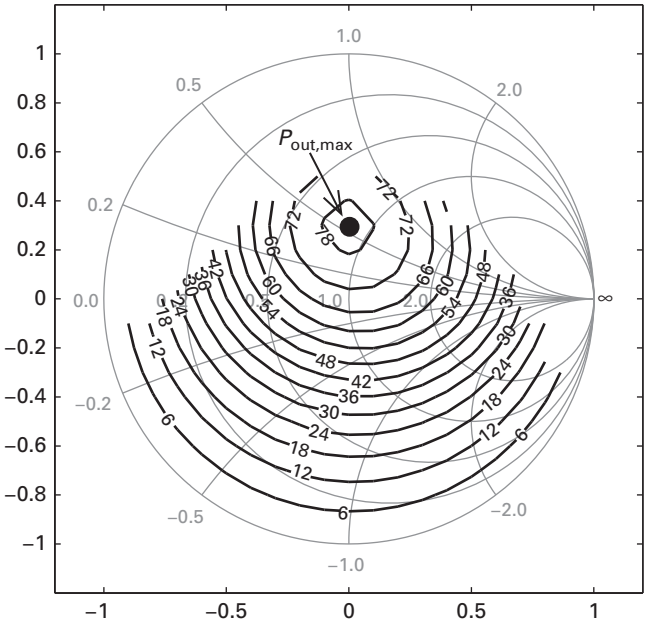


Figure 5.16 Output power contour obtained from RTALP at the fundamental frequency.

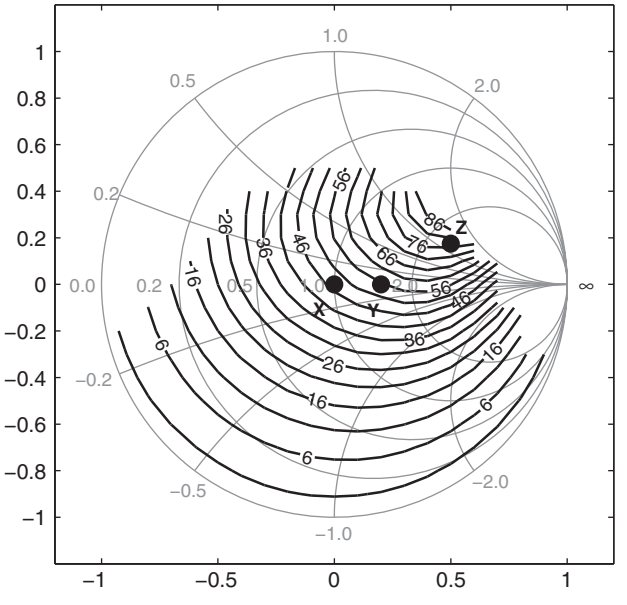


Figure 5.17 PAE contour obtained from RTALP at the fundamental frequency.

Figures 5.16 and 5.17. Various optimal load impedances can be selected in Figure 5.17 for achieving high PAE while keeping a large output power for this device. The optimal impedance corresponding to the case Z (marked by a black dot in Figure 5.17) achieves an efficiency of 80%. Other load impedances with higher PAE than that of case Z are

Table 5.3. Comparison of the ALP and RTALP at the fundamental frequency for the load impedance $\Gamma_L(\omega_0)$ at X, Y, and Z. (From [10] with permission, ©2006 IEEE.)

	PAE	(%)	Output power	(mW)
$\Gamma_L(\omega_0)$	RTALP	ALP	RTALP	ALP
Case X: 0	47.9	49.1	69.9	72.8
Case Y: $0.2\angle 0°$	55.4	57.6	67.1	72.4
Case Z: $0.53\angle 19.3°$	80.0	70.6	53.0	56.2

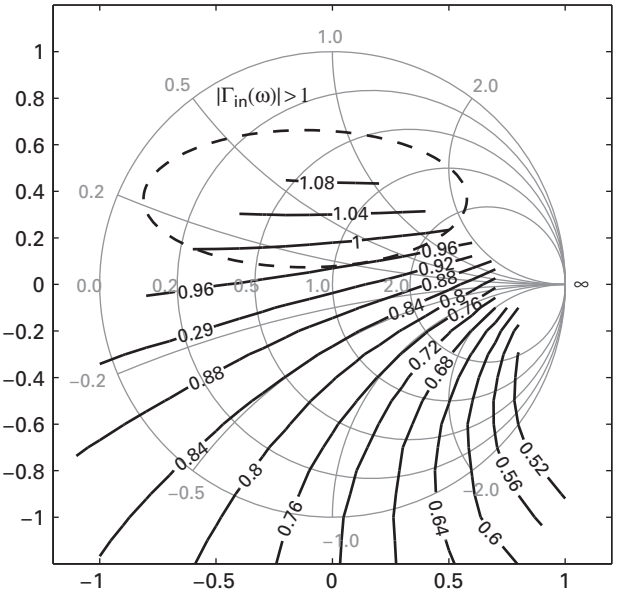


Figure 5.18 $|\Gamma_{in}(\omega_0)|$ contour plot at the fundamental. (From [10] with permission, ©2006 IEEE.)

observed in Figure 5.17. However, on referring to Figure 5.18, one can identify these termination impedances as potentially unstable because the corresponding $\Gamma_{in}(\omega_0)$ is greater than unity, indicating a negative impedance at the input of the transistor.

For each of the three load-impedance cases X, Y, and Z, the PAE and output power data extracted from RTALP and ALP correlate well, the largest difference being observed for the case Z when the efficiency is the largest. The differences observed in Table 5.3 are brought about by the memory effects excited by the modulation in RTALP.

On account of its high PAE and stability considerations, the load impedance located at Z is selected as the optimum choice for the rest of the class-F design.

Next RTALP measurements are conducted for the second-harmonic frequency for 18 power levels injected at the output, while maintaining at port 2 the same optimal reflection coefficient (case Z) at the fundamental frequency and a match termination for the third harmonic.

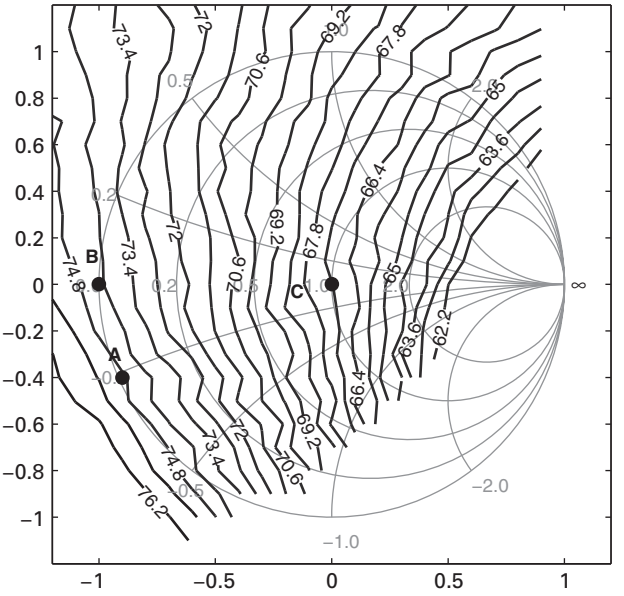


Figure 5.19 PAE contour obtained from RTALP at the second harmonic. (From [10] with permission, ©2006 IEEE.)

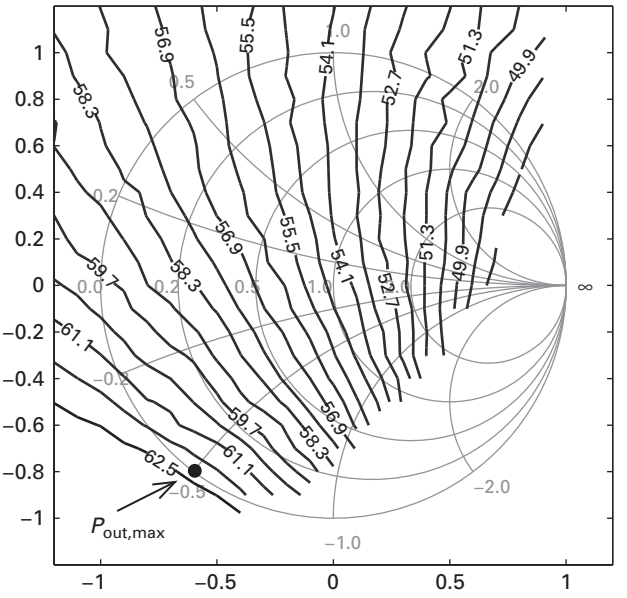


Figure 5.20 Output power contour obtained from RTALP at the second harmonic. (From [10] with permission, ©2006 IEEE.)

Both PAE and output power contours are shown in Figures 5.19 and 5.20. In agreement with both class-B and class-F PA design theory, the terminations corresponding to the highest PAE (case A in Figure 5.19) and the highest output power in Figure 5.20 are observed to be on the edge of the Smith chart and to approach an ideal

Table 5.4. Comparison of second-harmonic ALP and RTALP results for the load impedance $\Gamma_L(2\omega_0)$ at A, B, and C. (From [10] with permission, ©2006 IEEE.)

	PAE (%)		Output power (mW)	
$\Gamma_L(2\omega_0)$	RTALP	ALP	RTALP	ALP
Case A: $1\angle -156°$	74.95	72.03	61.15	57.7
Case B: $1\angle -180°$	74.18	71.82	59.79	57.6
Case C: 0	67.13	66.47	54.6	53.4

Table 5.5. Comparison of ALP results for the four dynamic loadlines considered in Figure 5.21 for a load impedance at X, Y, and Z. (From [10] with permission, ©2006 IEEE.)

Loadline	$\Gamma_L(\omega_0)$	$\Gamma_L(2\omega_0)$	PAE (%)	Output power (mW)
Case L1	$0.53\angle 19°$	$1\angle -156°$	72.0	57.7
Case L2	$0.2\angle 0°$	$1\angle -127°$	65.7	88.9
Case L3	0	$1\angle -156°$	58.1	96.5
Case L4	0	0	51.3	76.8

short. The highest PAE of 75% measured with RTALP is verified using ALP to be actually 72% when memory effects are eliminated, e.g. CW self-heating. The corresponding output power is 58 mW and the gain 17 dB.

Given that transistors are not perfect, a reactive impedance termination is indeed typically needed for the second harmonic. This situation has been analyzed in detail in the literature [1] [12], and a new class of circuits (class J) [1] has been proposed in order to address this practical situation. We shall see in the next section that the type of reactance needed depends on the device measured (packaged or on wafer) and the biasing circuit.

Note also that it is necessary to simultaneously optimize the fundamental and second-harmonic impedance terminations, but it is usually sufficient in practice to do a few iterative steps between the fundamental and the second-harmonic RTALP measurements to find the joint optimal terminations.

In Table 5.4, the PAE and output power for three typical cases A, B, and C corresponding, respectively, to the optimal, short-circuit, and match in Figure 5.19 are compared. The difference between RTALP and ALP is much less than that in Table 5.3, indicating that memory effects are more prominent at the fundamental frequency than at the second harmonic.

Further insight into the role of the second-harmonic termination is revealed in Table 5.5, which compares the PAE and output power obtained for different combinations of the fundamental and second-harmonic load conditions. Case L1 corresponds to the maximum PAE (the same as case A in Table 5.4). Case L3 corresponds to the maximum output power. Case L2 is a compromise between L1 and L3. Case L4 corresponds to the reference 50-Ω load condition.

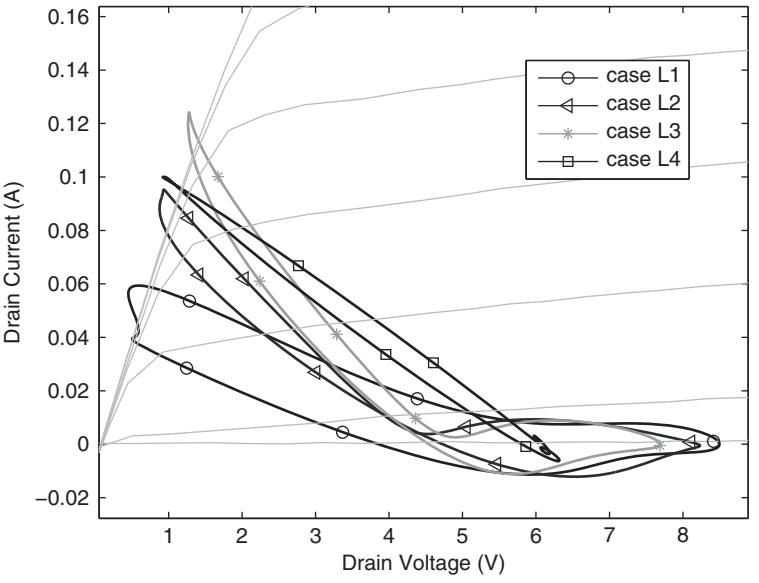


Figure 5.21 Loadlines associated with case L1 (maximum PAE), L3 (maximum power), L2 (hybrid), and L4 (matched load) impedance termination for the second harmonic. (From [13] with permission, ©2008 IEEE.)

Figure 5.21 shows the dynamic loadlines corresponding to cases L1, L2, L3, and L4. In case L4 a 50-Ω termination is used, and the device does not operate in class B, as indicated by the reduced drain-voltage swing. Cases L1, L2, and L3 feature about the same drain-voltage swing. They rely on a lossless reactive termination, which, like the shunt resonator in the ideal class-B device, allows the drain voltage to reach nearly twice the supply voltage V_{DD}. Case L3, for which the drain current in the dynamic loadline reaches the highest values, delivers the most output power. Case L1, for which the drain current remains small, achieves the highest PAE. The reason is that the knee voltage is reduced at lower drain currents. However, a further reduction in drain current will not improve the PAE, since the device gain drops when the device operates more in the threshold region.

Finally, RTALP measurements are conducted for the third-harmonic frequency for 21 power levels injected at the output, while maintaining at port 2 the optimal reflection coefficients of case Z at the fundamental frequency and case A-L1 at the second harmonic.

Both PAE and output power contours are shown in Figures 5.22 and 5.23. In agreement with class-F PA design theory, the terminations corresponding to the highest PAE (case A in Figure 5.19) and the high output power in Figure 5.20 are observed to be close to an open circuit and are on the edge of the Smith chart. The highest PAE of 80.96% is verified using ALP. This corresponds to an improvement of 8.9% over case A (best second-harmonic tuning). The corresponding output power and gain are 65 mW and 16.33 dB.

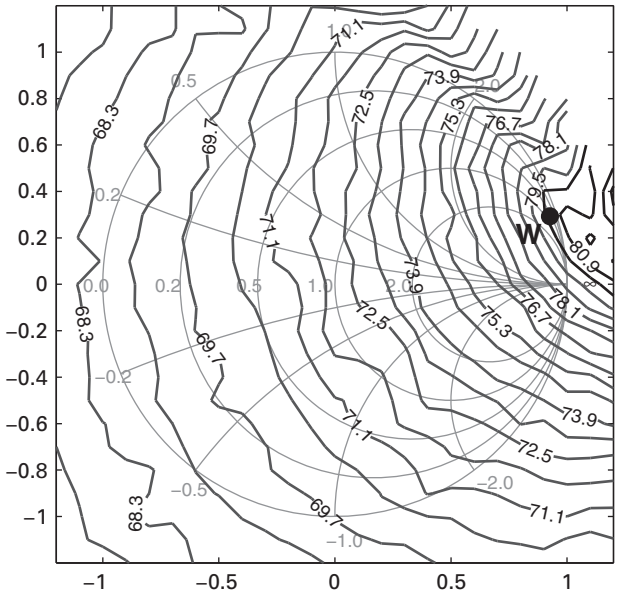


Figure 5.22 PAE contour obtained from RTALP at the third harmonic. (From [10] with permission, ©2006 IEEE.)

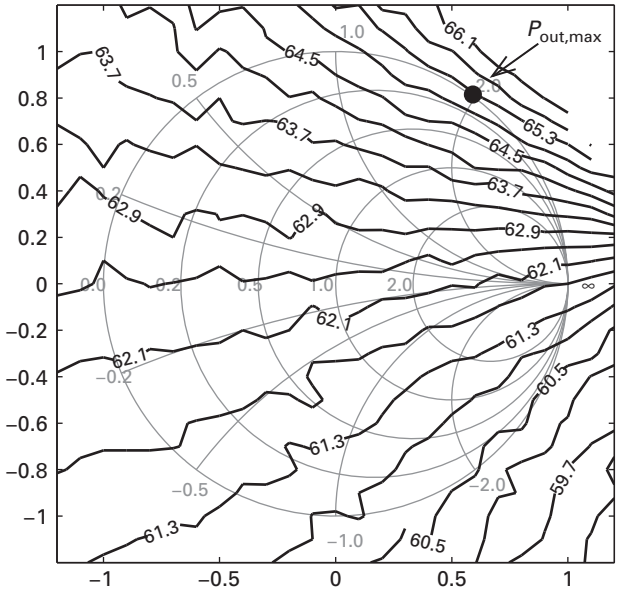


Figure 5.23 Output power contour obtained from RTALP at the third harmonic. (From [10] with permission, ©2006 IEEE.)

Despite the departure from an ideal open termination for the third harmonic, the RTALP experimental results obtained for an on-wafer GaN HEMT are consistent with the class-F PA theory, which indicates that highly reactive terminations for the transistor enable one to maximize the power efficiency.

5.4 Complete design cycle for a pHEMT amplifier

In the previous section we presented multi-harmonic RTALP results conducted on an on-wafer GaN HEMT. In this section we turn our attention toward the characterization of a packaged pHEMT device with RTALP for the subsequent design of a power-efficient amplifier at 2 GHz [13]. The device is biased around cutoff ($V_{DS} = 3$ V and $V_{GS} = 0.31$ V).

The amplifier layout shown in Figure 5.24 is modular so that the input-matching, transistor, and output-matching circuits can be tested individually, and also can mate directly with one another. This is ideal for metrology purposes, but obviously the PAE will be degraded by the loss of the four additional female and male SMA connectors used in the modular amplifier in addition to the input and output connectors. The bias tees are included with the transistor for convenience. Note, however, that this is not an optimal choice, insofar as the bias tees provide a fixed impedance termination for each of the harmonics, which constrains the PA optimization. It would have been preferable to use external broadband bias tees to allow optimization of the bias-tee location in the input- and output-matching networks.

The PAE contour plots obtained from the RTALP for the second harmonics are shown in Figure 5.25. A maximum PAE of 72% with an output power of 61 mW is obtained for the second-harmonic termination labeled A in Figure 5.25. As indicated in Table 5.6 this corresponds to a 6.6% increase in efficiency.

Various dynamic loadlines obtained from RTALP are shown (gray lines) in Figure 5.26, together with the optimal loadline (black line) retained. Maximum PAE is obtained when the drain currents achieve a large excursion in the region of negative current. This corresponds to the packaged pHEMT releasing stored reactive energy. The corresponding drain current and voltage waveforms are shown in Figure 5.27(a), together with the instantaneous power dissipated $v_{DS}(t) \times i_{DS}(t)$ in Figure 5.27(b). The PAE is maximized when the difference between the positive and negative areas is the smallest possible fraction of the dissipated DC power. This results directly from the definition of PAE,

$$\mathrm{PAE} = \frac{P_{RF} - P_{in}}{P_{DC}} = 1 - \frac{P_{diss}}{P_{DC}}, \tag{5.7}$$

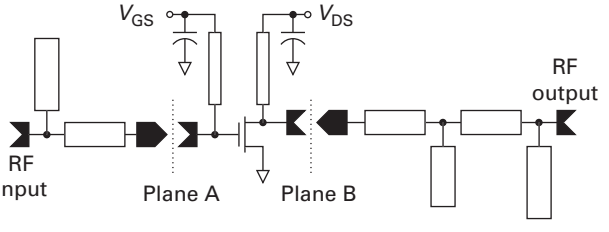


Figure 5.24 Simplified schematic diagram of the designed pHEMT amplifier including bias tees. The genders for the SMA connectors are specified. (From [13] with permission, ©2008 IEEE.)

Table 5.6. Comparison of ALP and RTALP for the load impedance at A, B, and C for the pHEMT. (From [13] with permission, ©2008 IEEE.)

	PAE (%)		Output power (mW)	
$\Gamma_L(\omega_0)$	RTALP	ALP	RTALP	ALP
Case A: $0.6 + j0.7$	77.0	71.9	60.3	60.8
Case B: short-circuit $1\angle 180°$	62.7	63.5	52.1	51.4
Case C: 0	60.4	61.9	46.9	47.1

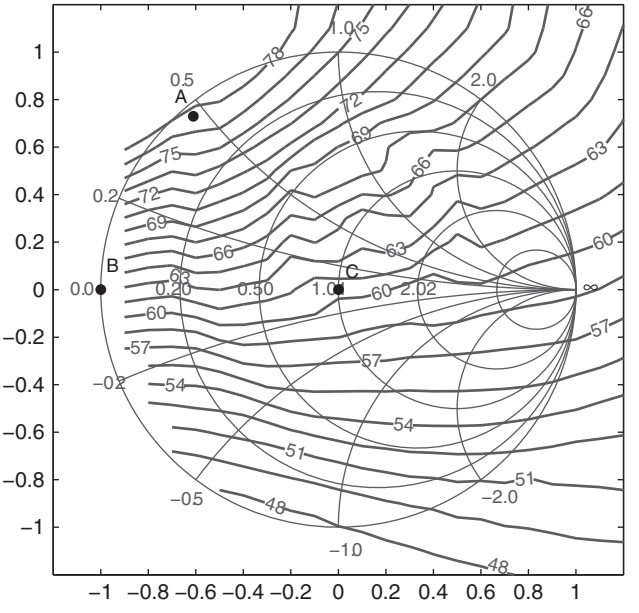


Figure 5.25 PAE contour obtained from RTALP at the second harmonic. (From [13] with permission, ©2008 IEEE.)

expressed in terms of the total power dissipated P_{diss}:

$$P_{\text{diss}} = \frac{1}{T}\int_0^T [v_{\text{GS}}(t)i_{\text{G}}(t) + v_{\text{DS}}(t)i_{\text{D}}(t)]dt,$$

using the total power conservation $P_{\text{DC}} + P_{\text{in}} = P_{\text{RF}} + P_{\text{diss}}$.

The fabricated amplifier with bias tees (Figure 5.24) is measured with the reference planes deembedded to planes A and B to obtain the voltage and current waveforms near the transistor terminals. From the power-sweep measurement shown in Figure 5.28 we can observe a PAE of 68.5% with output power 18 dBm and gain 16 dB at 2 GHz. The

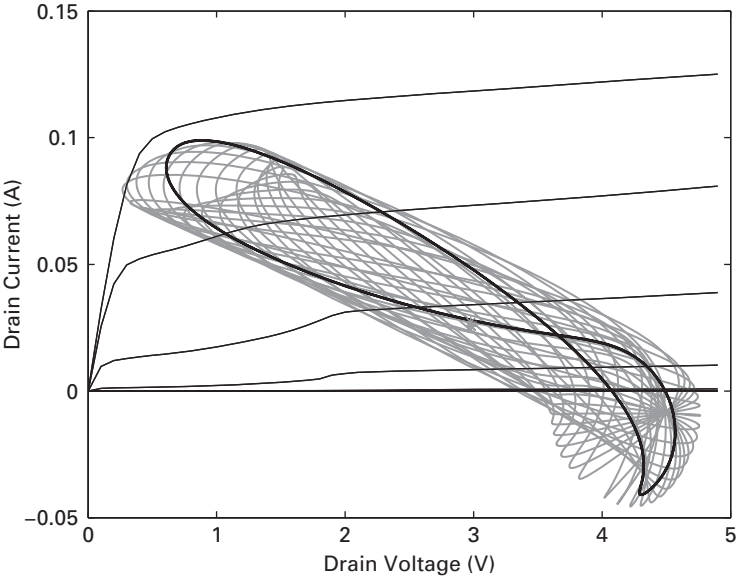


Figure 5.26 Dynamic loadlines for a pHEMT obtained from RTALP (gray lines) with the optimal loadline (black line) selected for maximum PAE. (From [13] with permission, ©2008 IEEE.)

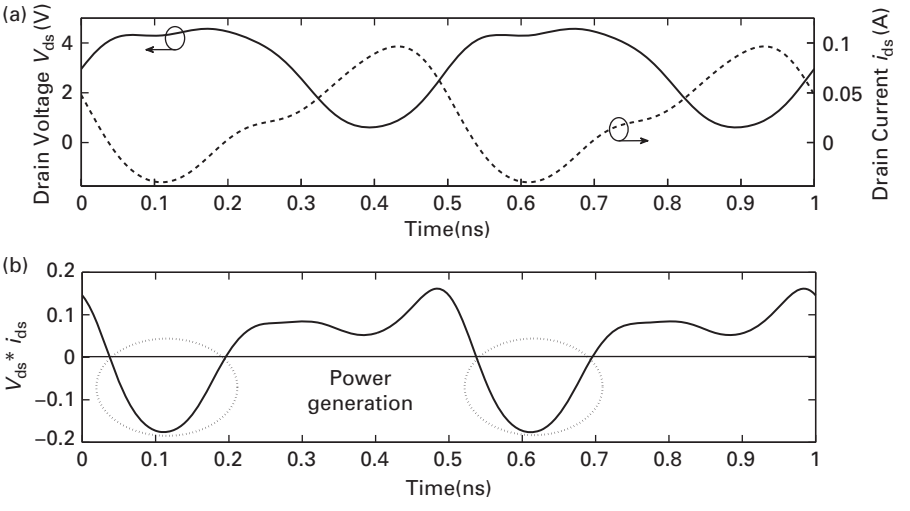


Figure 5.27 Drain current and voltage waveforms (top) and instantaneous power dissipation at the drain (bottom). (From [13] with permission, ©2008 IEEE.)

PAE and output-power results are very close to the RTALP findings in Table 5.6. Below the 1-dB compression point, this PA achieves a PAE of up to 74.6% with output power 18.6 dBm. These results indicate the usefulness of RTALP to facilitate the design of PAs while greatly shortening the design cycle.

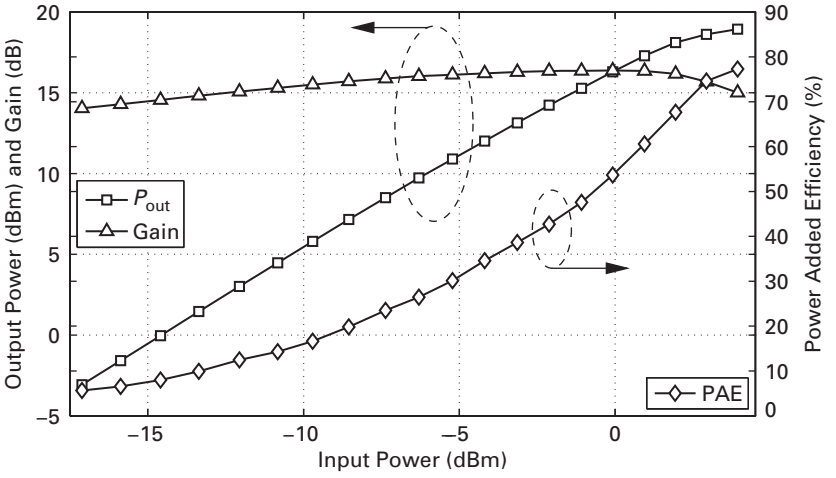


Figure 5.28 Output power, gain, and PAE versus input power for the pHEMT amplifier fabricated. (From [13] with permission, ©2008 IEEE.)

5.5 RTALP of PAs for pulsed *I–V* pulsed-RF class-B operation

The ALP and RTALP techniques described in the previous section can also be applied to the design of PAs optimized for pulsed *I–V* pulsed-RF operation. In such a case the ALP testbed shown in Figure 5.29 is simply modified by (1) adding multi-harmonic pulsed RF sources at the output port and (2) combining the pulsed biasing with the pulsed RF signal using broadband bias tees.

Two modes of data recording are possible when using NVNAs for pulsed RF signals. For sampler-based NVNAs (LSNA, SWAP) multiple recordings can be used to record the pulsed RF signal while performing the RTALP measurements. Indeed, while using the multiple-recording mode, it is possible to acquire modulated signals by sampling the signal at the time of interest within the pulse duration. The theory for this mode of operation was presented in Section 2.9.2 [14]. Successful RTALP has been demonstrated using this approach. Alternatively, a sampler-based NVNA can be used to acquire the full spectra of the pulsed signals (see Section 2.8). In such a case, only ALP measurements have been attempted [15], and some of the results obtained are reported below.

The validity of the pulsed I–V pulsed-RF technique for class-B operation is verified with the measurement of a device that does not suffer much from memory effects. A 65-nm MOSFET is used as such a reference device. As shown in Figure 5.30, the quasi-orthogonal voltage and current waveforms measured under DC *I–V* and CW-RF excitation (dashed lines) are seen to compare very well with the ones obtained using pulsed *I–V* biasing and pulsed-RF excitations (plain lines). The pulsed *I–V* duration is 1 μs with a duty rate of 1%. The pulsed-RF duration is 0.33 μs with a duty rate of 0.33%. Note that class-B operation was achieved using an adaptive technique for the second harmonic [16].

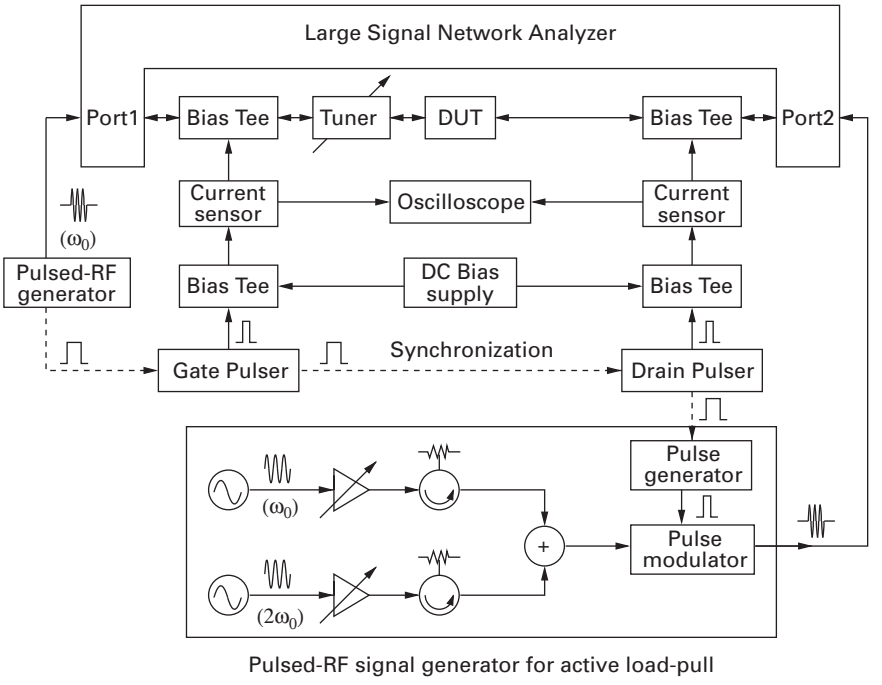


Figure 5.29 Modified active load-pull testbed for pulsed *I–V* pulsed-RF measurements. (From [15] with permission, ©2009 IEEE.)

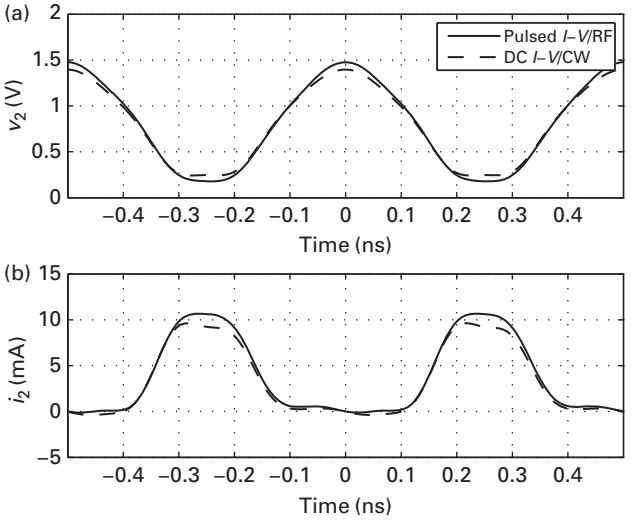


Figure 5.30 Optimized output voltage (a) and current (b) waveforms for the pulsed *I–V* pulsed-RF operation of a 65-nm MOSFET. (From [15] with permission, ©2009 IEEE.)

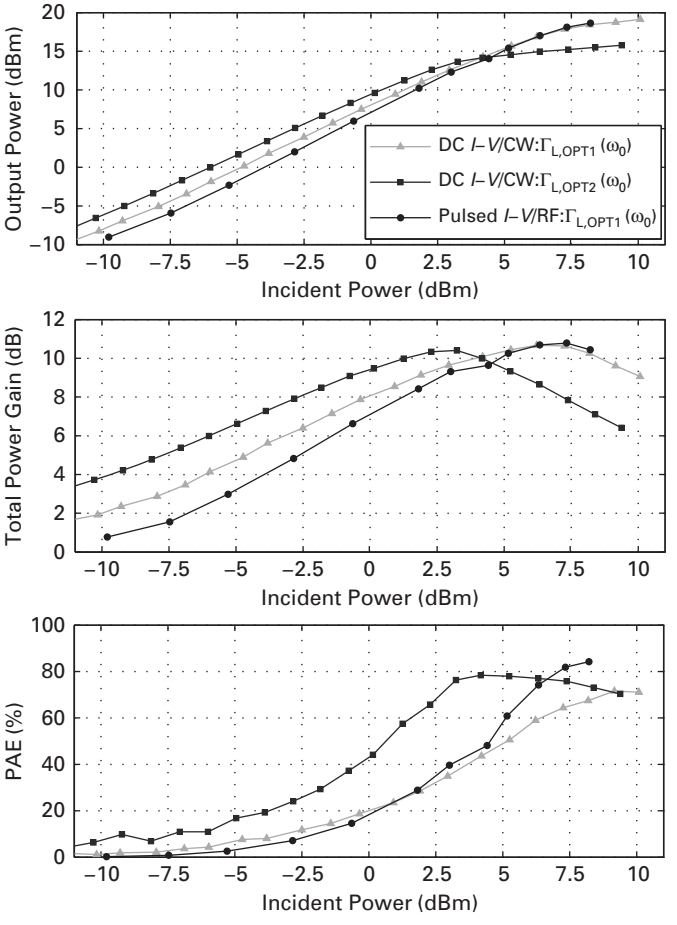


Figure 5.31 Output power, total power gain, and PAE measured for three different load terminations at the second harmonic. (From [15] with permission, ©2009 IEEE.)

The pulsed *I–V* pulsed-RF ALP system described above is then applied to a passivated GaN HEMT device exhibiting some residual memory effects. The goal is to compare the pulsed *I–V* pulsed-RF performance of the transistor operated in class B while using the optimal second-harmonic termination obtained from (1) CW signal load-pull and (2) pulsed *I–V* pulsed-RF load-pull. Both the terminations yielding optimal PAE and optimal output power for the CW load-pull are considered.

The output power, total power gain, and PAE are shown in Figure 5.31. It is verified in this design example that only the optimal termination obtained from the pulsed *I–V* pulsed-RF load-pull is able to simultaneously achieve a high PAE with a high gain and output power. This suggests that memory effects, e.g. self-heating, contribute to the segregation of the loading conditions for optimal output power and PAE. Note also that the gain of this class-B/C amplifier is strongly nonlinear. Figure 5.32 shows the output voltage and current under pulsed *I–V* pulsed-RF operation for increasing input power levels, and Figure 5.33 shows the associated dynamic loadlines. Clearly the optimal

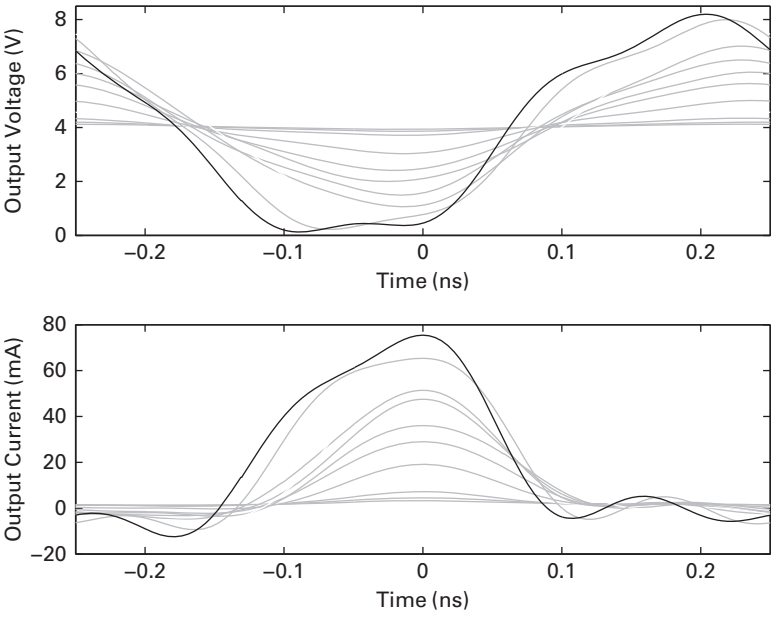


Figure 5.32 Output voltage and current waveforms of the AlGaN/GaN HEMT on SiC in the pulsed *I–V* pulsed-RF measurements for increasing input power. (From [15] with permission, ©2009 IEEE.)

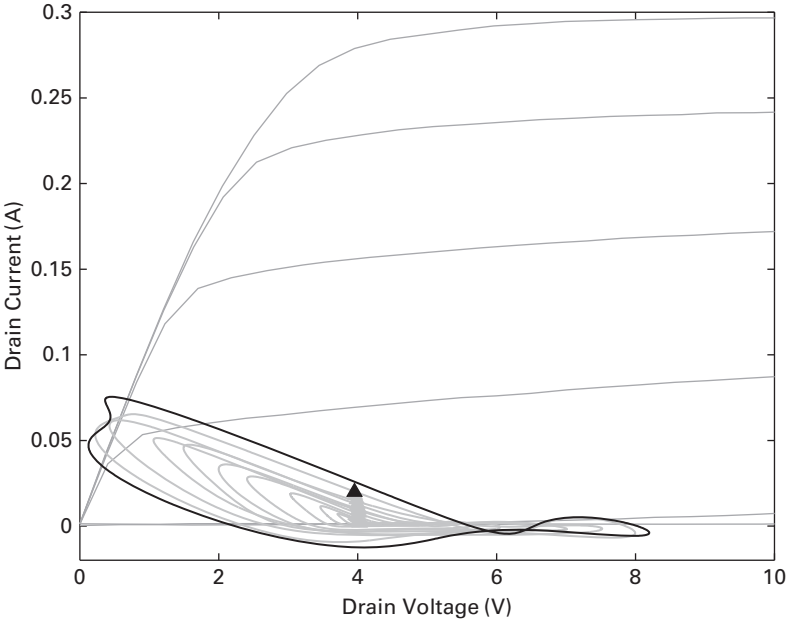


Figure 5.33 Associated dynamic loadlines for the various waveforms shown in Figure 5.32. (From [15] with permission, ©2009 IEEE.)

operation (orthogonal current and voltage waveforms) yielding 80% PAE is achieved only at maximum input power. Also the optimal PAE obtained again corresponds to operation at low maximum drain currents so as to keep the knee voltage small. As is characteristic of class-C amplifiers, the amplifier requires a minimum input power to turn on. This type of efficient but nonlinear amplifier is typically used in radar systems or as the peaking amplifier in Doherty amplifiers.

5.6 P1dB contour plot

Note that, in the RTALP discussed in this chapter, the gate-voltage swing was set to an appropriately large value and kept the same in all subsequent measurements. Linear operation is also of great importance in amplifier design. P1dB contour plots then provide a useful figure of merit for maximizing the output power while retaining a linear operation.

P1dB contour plots can be generated also using RTALP by sweeping the input power in addition to the output power sweep. For a given input power level, only the RTALP data (PAE, gain, output power) occurring at load impedances at which the gain drops by 1 dB relative to the small-signal gain are then retained for each input power. The accumulated P1dB RTALP data retained for all input power levels can then be used to generate a P1dB power contour plot as well as P1dB gain and P1dB PAE contour plots. An example of a P1dB power contour plot obtained is shown in Figure 5.34 for a GaN

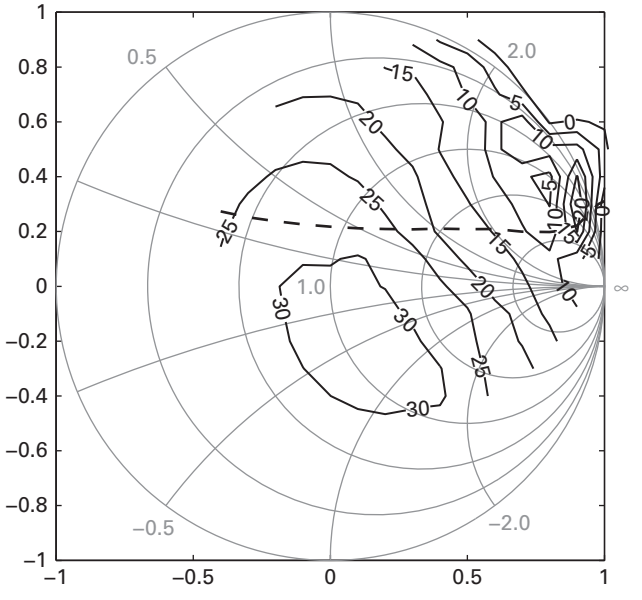


Figure 5.34 A P1dB output power contour plot for pulsed-RF excitation of duty rate 0.33% and duration 0.33 μs. The dashed line separates the stable (bottom) and potentially unstable (top) regions. The unit of the label is mW. (Measured by Youngseo Ko at Ohio State University.)

HEMT operated in class A with pulsed-RF excitation of duty rate 0.33% and duration 0.33 μs. The P1dB power contour plot was generated on the basis of data acquired with pulsed-RF RTALP using a modulation of $\Delta f = 200$ kHz. Pulsed-RF RTALP relies on the method of multiple recording modified as described in Section 2.9.2 to work with signals that are both pulsed and modulated.

5.7 Class-E PA operation

To conclude this chapter, results obtained on a class-E amplifier design with a TSMC 0.18 μm CMOS transistor are presented. The operating frequency is 3.5 GHz. Five harmonics are acquired. The device is biased with $V_{GS} = 0.4$ V and $V_{DS} = 0.6$ V. The drain current is 0.5 mA. A larger DC drain voltage would have been preferred, but the measurement setup used was limited by the RF power available from the output amplifier of the signal source. The transistor features a shunt capacitor of 5.1 pF. A corporate tree is used for the device layout for a total chip area of 0.55 mm × 0.60 mm.

Real-time active load-pull (RTALP) measurements are first conducted at the fundamental using an open circuit for the second harmonic ($\Gamma_L(2\omega_0) = 1$). The higher harmonics are terminated by a matched load. The maximum PAE and output power are predicted by RTALP to be 55.8% and 8.85 mW, respectively, at the optimal load impedance $\Gamma_L(\omega_0) = 0.922\angle 167°$. Since the RTALP results may suffer from memory effects, the actual performance of the device is verified using active load-pull (ALP). A PAE of 60.4% and an output power of 8.9 mW (9.5 dBm) are obtained using the optimal termination determined by RTALP. Optimization of the second harmonic using RTALP was also performed. It was found that for this device the impedance termination $\Gamma_L(2\omega_0)$ had only a small impact on the PAE and gain, and the ideal open-circuit termination was retained.

The drain current and voltage waveforms are shown in Figures 5.35(a) and (b) for different input power levels. The total current (plain line) is quasi-sinusoidal due to the second-harmonic open termination. Since the added capacitance is known, the capacitor current can be reconstituted (dotted lines) from the drain voltage and subtracted from the measured current to yield the drain current of the transistor (dashed line). As expected, the drain current and voltage waveforms are found to exhibit reduced overlap after the capacitor current has been accounted for. However, it is seen that the switching action of the transistor is imperfect but the capacitor current does take over once the drain current has decreased.

Figure 5.35 reveals that the drain voltage waveforms do not turn off completely. This is due to the knee voltage. This is further revealed by the dynamic loadlines plotted versus gate voltage and drain voltage in Figures 5.36(a) and (b). The large negative current observed in part of the cycle is associated with the shunt capacitor placed across the drain. The small residual negative drain current observed is associated with the transistor drain-to-source capacitance. According to the dynamic loadline of Figure 5.36(b) the

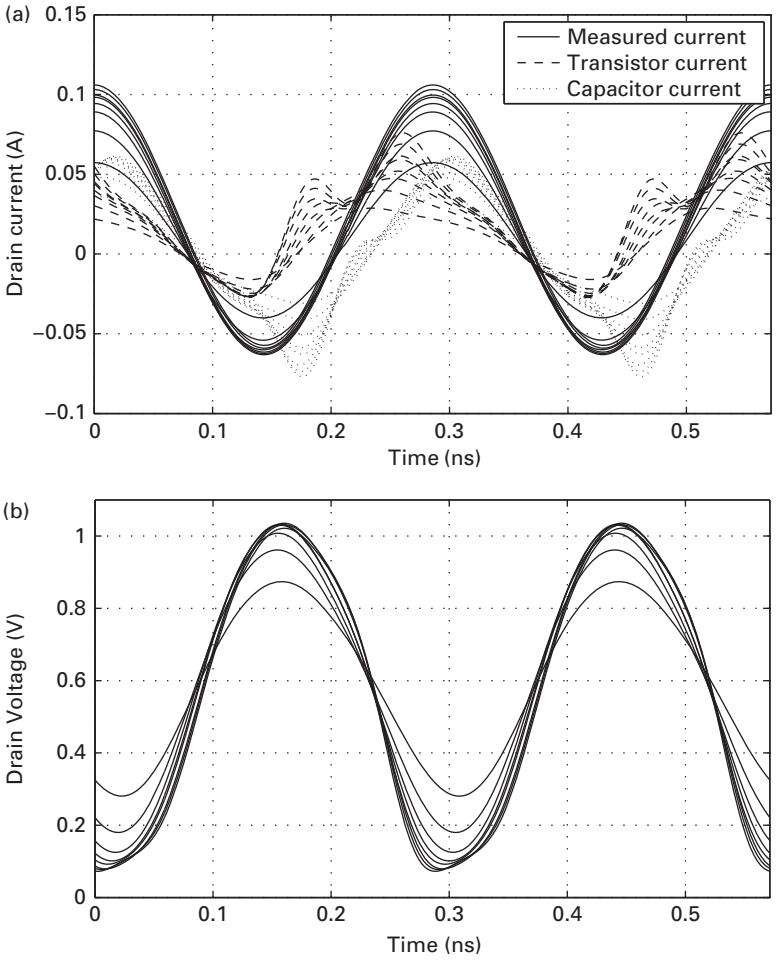


Figure 5.35 Waveforms for the measured, transistor, and capacitor output currents (a) and drain voltage (b) for different input power levels.

knee of the CMOS transistor is about 0.2 V when the transistor is on. This is relatively large compared with the DC drain bias voltage of 0.6 V. The associated on resistance is larger than 2 Ω. The non-ideal switching behavior of the transistor is thus responsible for the reduced PAE (60.4%) compared with that predicted by the ideal class-E theory.

An improved performance could also have been obtained by driving the input with a square-wave signal, instead of a sinusoidal input. Alternatively, the input power was swept and the resulting output power P_{out} and PAE versus input power are shown in Figure 5.37. A maximum PAE of 66.5% is obtained for an input power of 0.5 mW, yielding an output power of 6.5 mW (8.13 dBm) and a gain of 13.4 dB at 3.5 GHz. For comparison, a class-F amplifier realized with the same transistor without the shunt

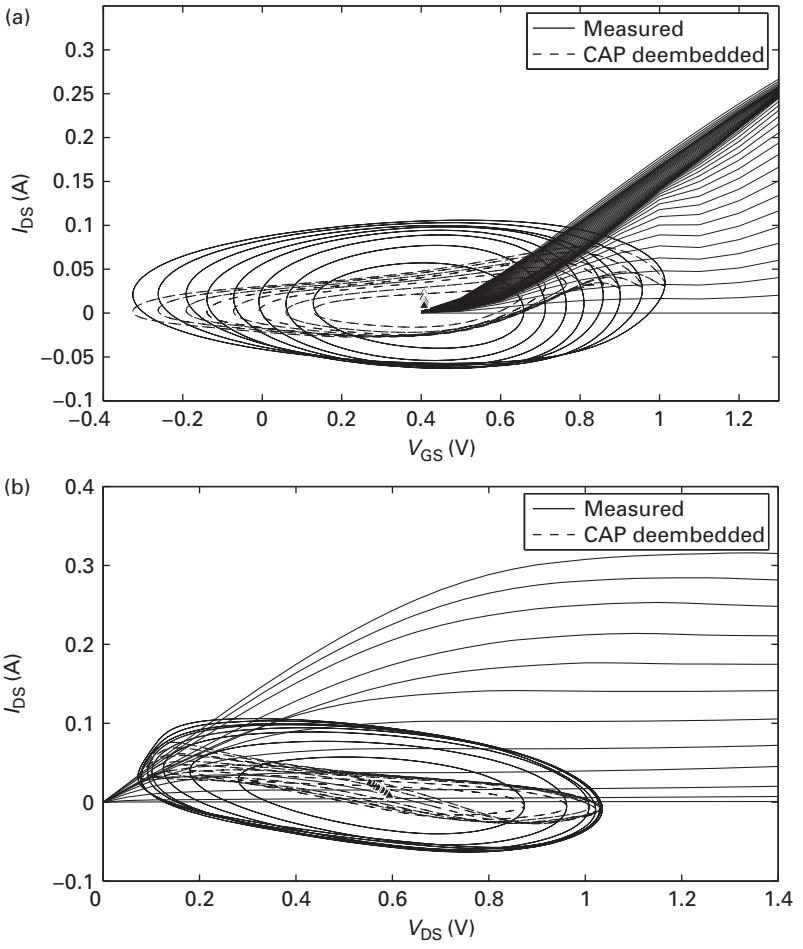


Figure 5.36 Loadline versus gate and drain voltages for class-E operation. The superposed DC *I–V* characteristics are from 0 to 1.5 V in steps of 0.1 V.

capacitor yielded a PAE of 71%, an output power of 6.4 dBm, and a gain of 14 dB at 3.5 GHz while using the same low drain voltage of 0.6 V.

These measurements clearly indicate that the performance and waveforms predicted for class-E amplifier operation are difficult to achieve at low bias drain voltages from a real transistor exhibiting a large on resistance and knee voltage. The use of a larger drain voltage will improve the PAE performance of the circuit tested. Much better performance can be achieved using a technology sustaining large drain voltages without device breakdown, such as the GaN HEMT technology.

Overall in this section our intention was to demonstrate that, as for class F, the use of RTALP with an NVNA can help search for the optimal class-E performance which can be achieved with a specific technology.

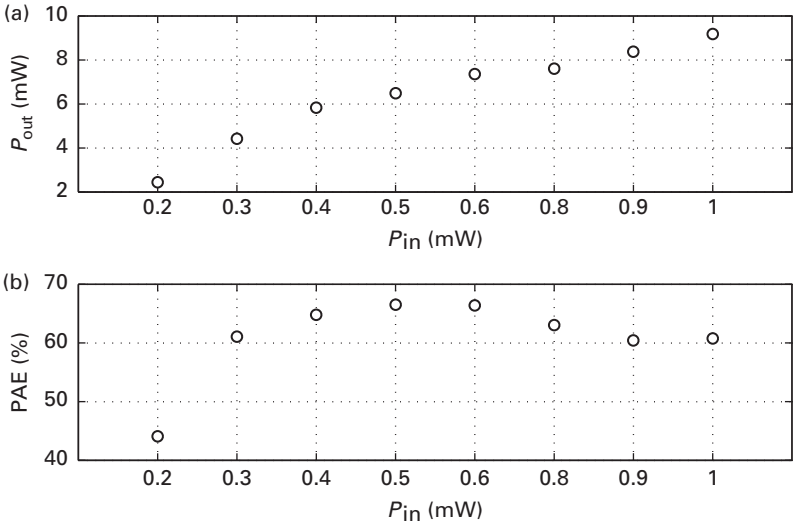


Figure 5.37 Output power (a) and PAE (b) of a class-E amplifier at 3.5 GHz.

References

[1] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*, 2nd edition, Artech House, 2006.

[2] N. O. Skal and A. D. Sokal, "Class E, a new class of high-efficiency tuned single-ended switching power amplifiers," *IEEE Journal of Solid State Circuits*, Vol. 10, No. 6, pp. 168–176, June 1975.

[3] F. H. Raab, "Idealized operation of the class E tuned power amplifier," *IEEE Transactions on Circuits and Systems*, Vol. 24, No. 12, pp. 725–735, Dec. 1977.

[4] F. H. Raab, "Class-E, class-C, and class-F power amplifiers based upon a finite number of harmonics," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 49, No. 8, pp. 1462–1468, Aug. 2001.

[5] Y. Takayama, "A new load-pull characterization method for microwave power transistors," in *IEEE MTT-S International Microwave Symposium Digest*, pp. 218–220, 1976.

[6] G. P. Bava, U. Pisani, and V. Pozzolo, "Active load technique for load-pull characterization at microwave frequencies," *Electronic Letters*, Vol. 18, No. 2, pp. 178–180, Feb. 1982.

[7] D. Barataud, F. Blache, A. Mallet, P. P. Bouysse, J.-M. Nebus, J. P. Villotte, J. Obregon, J. Verspecht, and P. Auxemery, "Measurement and control of current/voltage waveforms of microwave transistors using a harmonic load-pull system for the optimum design of high efficiency power amplifiers," *IEEE Transactions on Instrumentation and Measurement*, Vol. 48, No. 8, pp. 835–842, Aug. 1999.

[8] J. Benedikt, R. Gaddi, P. J. Tasker, and M. Goss, "High-power time-domain measurement system with active harmonic load-pull for high-efficiency base-station amplifier design," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 48, No. 12, pp. 2617–2624, Dec. 2000.

[9] F. Verbeyst and M. Vanden Bossche, "Real-time and optimal PA characterization speeds up PA design," in *34th European Microwave Conference Digest*, Amsterdam, pp. 431–434, 2004.

[10] X. Cui, S. J. Doo, P. Roblin, G. H. Jessen, R. G. Rojas, and J. Strahler, "Real-time active load-pull of the 2nd & 3rd harmonics for interactive design of non-linear power amplifiers," in *ARFTG 68th Conference Digest*, Colorado, 2006.

[11] P. Roblin, S. J. Doo, X. Cui, G. H. Jessen, D. Chaillot, and J. Strahler, "New ultra-fast real-time active load-pull measurements for high speed RF power amplifier design," in *IEEE MTT-S International Microwave Symposium*, Honolulu, HI, 2007.

[12] P. Colantonio, F. Giannini, G. Leuzzi, and E. Limiti, "High efficiency low-voltage power amplifier design by second harmonic manipulation," *International Journal of RF and Microwave Computer Aided Engineering*, Vol. 10, No. 1, pp. 19–32, Jan. 2000.

[13] X. Cui, S. J. Doo, P. Roblin, J. Strahler, and R. G. Rojas-Teran, "High efficiency RF power amplifier designed with harmonic real-time active load-pull," *IEEE Microwave and Wireless Components Letters*, Vol. 18, No. 4, pp. 266–268, April 2008.

[14] F. De Groote, P. Roblin, Y. Ko, and C. Yang, "Pulsed multi-tone measurements for time domain load pull characterizations of power transistors," in *IEEE MTT-S International Microwave Symposium*, Boston, MA, 2009.

[15] S. J. Doo, P. Roblin, V. Balasubramanian, R. Taylor, K. Dandu, J. Strahler, G. H., Jessen, J.-P. Teyssier, "Pulsed active load-pull measurements for the design of high-efficiency class-B RF power amplifiers with GaN HEMTs," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 57, No. 4, pp. 881–889, April 2009.

[16] S. J. Doo, P. Roblin, V. Balasubramanian, R. Taylor, K. Dandu, G. H. Jessen, and R. Rojas, "Adaptive second harmonic active load for pulsed-*IV*/RF class-B operation," in *ARFTG 70th Conference*, Tempe, AZ, 2007.

6 Behavioral modeling[1]

6.1 Behavioral model for SISO and MIMO systems

In this chapter we will discuss various new techniques for the behavioral modeling of devices characterized by NVNA and VSA measurements. The reader is referred to the companion series book [1] for a comprehensive review of the field of behavioral modeling.

In our presentation we will focus on the development of single-input single-output (SISO) models. As shown in Figure 6.1(a) SISO models can be directly applied to single-port (two terminals) nonlinear loads or diodes. An example is the nonlinear negative resistance of an oscillator, which will be discussed in Chapter 7.

For the modeling of power amplifiers that are two-port devices (four terminals) a SISO model as shown in Figure 6.1(a) can be applied if we assume that the input and output ports are matched. That way, none of the reflected waves at the input $b_1(p\omega)$ from the device is converted into incident waves $a_1(p\omega)$ at port 1 and similarly none of the transmitted waves $b_2(p\omega)$ is converted into incident waves $a_2(p\omega)$ at port 2. A SISO model can also be applied to three-port modulator systems under similar approximations, as we shall see in Chapter 8.

The matching of the generator impedance and output load impedance to the characteristic impedance (typically 50 Ω) can be improved in practical systems by inserting attenuators or isolators at the input and output ports. Obviously, in the real world no device is perfect and some small reflections will still take place. This does not, however, invalidate the SISO model but limits it to work with the particular source and load impedances for which it was extracted. This points to the fact that modeling the PA combined with its filters, isolators, attenuators or driver will improve the portability of the model. Note also that if the amplifier can be assumed, to a high accuracy, to be unilateral (all $b_1(p\omega_0)$ independent of all $a_2(p\omega_0)$) then we can use a different SISO system to model separately both the reflected waves $b_1(p\omega_0)$ and the transmitted waves $b_2(p\omega_0)$ (see Figure 6.1(b)) and handle generators with arbitrary input impedances $\Gamma_G(p\omega_0)$. For harmonic excitations of such a two-port circuit we then have the following set of transcendental equations:

$$
\begin{aligned}
a_1(p\omega_0) &= \Gamma_G b_1(p\omega_0) + b_G(p\omega_0),\\
b_1(p\omega_0) &= F_{1,p}[a_1(\omega_0), a_1(2\omega_0), \dots],\\
b_2(p\omega_0) &= F_{2,p}[a_1(\omega_0), a_1(2\omega_0), \dots],
\end{aligned}
$$

[1] Research collaboration with Dominique Chaillot, Xi Yang, and Inwon Suh is gratefully acknowledged.

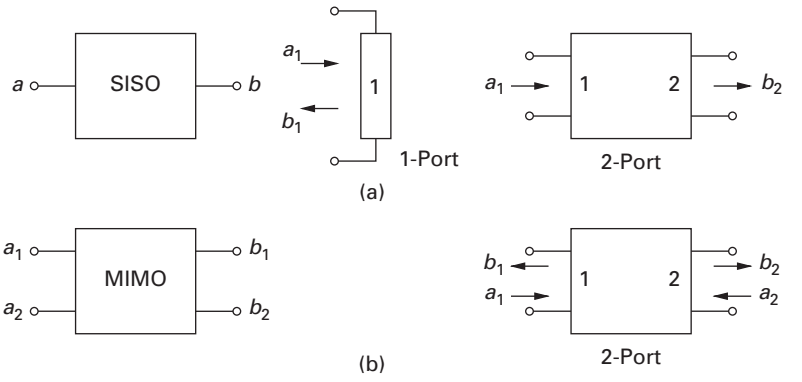


Figure 6.1 SISO (a) and MIMO (b) models for 1-port and 2-port circuits.

where the functions $F_{1,p}$ and $F_{2,p}$ are the describing functions [2] for the reflected waves at port 1 and port 2, respectively. The modeling is now that of a single-input multiple-outputs (SIMO) system, with the single input being the generator signals $b_G(p\omega_0)$.

In the general bilateral case, however, the describing functions for a two-port device will be dependent on both $a_1(p\omega_0)$ and $a_2(p\omega_0)$:

$$\begin{aligned} b_1(p\omega_0) &= F_{1,p}[a_1(\omega_0), a_1(2\omega_0), \ldots, a_2(\omega_0), a_2(2\omega_0), \ldots], \\ b_2(p\omega_0) &= F_{2,p}[a_1(\omega_0), a_1(2\omega_0), \ldots, a_2(\omega_0), a_2(2\omega_0), \ldots]. \end{aligned} \tag{6.1}$$

6.2 Volterra modeling

The most rigorous theory for including memory effects in nonlinear systems is the Volterra formalism [3] [4] [5] [6]. In that formalism the system is described by Volterra kernels of various orders. For example, for a SISO system/circuit, a third-order system is represented by three kernels h_1, h_2, and h_3 (multidimensional impulsed response), with an expansion of the form

$$\begin{aligned} y(t) &= y_1(t) + y_2(t) + y_3(t), \\ y_1(t) &= \int_{-\infty}^{\infty} h_1(\tau_1)x(t-\tau_1)d\tau_1, \\ y_2(t) &= \iint_{-\infty}^{\infty} h_2(\tau_1,\tau_2)x(t-\tau_1)x(t-\tau_2)d\tau_1\, d\tau_2, \\ y_3(t) &= \iiint_{-\infty}^{\infty} h_3(\tau_1,\tau_2,\tau_3)x(t-\tau_1)x(t-\tau_2)x(t-\tau_3)d\tau_1\, d\tau_2\, d\tau_3. \end{aligned} \tag{6.2}$$

The extension of the Volterra third-order model to higher order is easily guessed. The problem with the Volterra series representation is that the modeling of hard nonlinearities calls for the inclusion of very-high-order terms and the convergence of such a series is usually slow. Further, the Volterra model does not degrade graciously outside its range

of extraction. However, in some circumstances we can find a remedy to this problem, since, as we shall see, the functional representations are not necessarily limited to power series.

The Volterra expansion for high-order nonlinear systems becomes rather complex. The extension of the Volterra formalism to multiple inputs and multiple outputs, which is described in [3] [5], also increases the model complexity. However, as we shall discuss in the next section, a simpler picture emerges for the Volterra series if we limit our analysis to periodic input signals with a finite number of harmonics, and we switch to the frequency domain.

Limiting oneself to discrete tones might appear to be a severe limitation in today's world of digital modulation. However, it is possible to generalize the discrete-tone results obtained if the signals are assumed to be modulated with a sufficiently small bandwidth. Memory polynomials [7] [8] can then be used to handle wider modulation bandwidth. The resulting model is then both time- and frequency-selective.

6.2.1 Volterra algorithm

The Volterra series is a well-known approach to represent nonlinear circuits in the time and frequency domains. In the frequency domain we shall see that it admits a simple representation for a finite number of harmonic tones that is valid up to infinite order. For clarity we start by presenting the results obtained for from one up to four harmonic excitations before describing the general methodology used to derive these Volterra expansions.

Let us first consider injecting a signal consisting of a single incident harmonic wave $a(\omega_0)$ on a nonlinear SISO circuit with infinite order. All other incident harmonics $a(n\omega_0)$ with $n > 1$ are assumed to be negligible and are set to zero. The system being of infinite order, we expect that an infinite number of harmonics will be generated for the reflected waves. Let us consider only the first five reflected waves, $b(\omega_0)$, $b(2\omega_0)$, $b(3\omega_0)$, $b(4\omega_0)$, and $b(5\omega_0)$. It can be verified that they can then be represented up to infinite order using the following Volterra series expansion:

$$b(0) = V_{0,1}(\mathbf{x}_1), \tag{6.3}$$

$$b(\omega_0) = a(\omega_0) \cdot V_{1,1}(\mathbf{x}_1), \tag{6.4}$$

$$b(2\omega_0) = a^2(\omega_0) \cdot V_{2,1}(\mathbf{x}_1), \tag{6.5}$$

$$b(3\omega_0) = a^3(\omega_0) \cdot V_{3,1}(\mathbf{x}_1), \tag{6.6}$$

$$b(4\omega_0) = a^4(\omega_0) \cdot V_{4,1}(\mathbf{x}_1), \tag{6.7}$$

$$b(5\omega_0) = a^5(\omega_0) \cdot V_{5,1}(\mathbf{x}_1). \tag{6.8}$$

For a one-port device we will generally define $b(0)$ to be the DC-operation current $b(0) = I_0$. The functions $V_{i,1}(\mathbf{x}_1)$ in Equations (6.3)–(6.6) are found to be functionally dependent on $\mathbf{x}_1 = [x_{1,0}, x_{1,1}]$, which features two DC terms: the supply voltage $x_{0,1} = V_0 = a(0)$ and $x_{1,1} = |a(\omega_0)|^2$. Note that multiple voltages or current supplies could be introduced if needed to describe the biasing of the nonlinear device under consideration. Also *self-biasing* and any other dependent internal node voltages or temperatures that

affect the device operation are automatically accounted for by all the functions $V_{i,1}(\mathbf{x}_1)$. An example of self-biasing will be given in Section 6.2.3.

Further, note that the selection of the independent variable $a(n\omega)$ and the control variable $b(n\omega)$ is really up to the user. In Section 6.2.3 we will find it easier to select voltages and currents for $a(n\omega)$ and $b(n\omega)$ in the derivation of an analytic model of an ideal diode. However, in the RF measurement context such as in Section 6.2.5, finite-impedance (e.g. 50 Ω) generators are used in practice at RF. It is then preferable to use incident wave and reflected wave for a and b while keeping voltage(s) V_0 for the DC bias which is usually provided by a very-low-impedance power supply.

Let us now consider injecting a signal consisting of two incident harmonic waves $a(\omega_0)$ and $a(2\omega_0)$ on the nonlinear impedance of a SISO circuit with infinite order. All other incident harmonics $a(n\omega_0)$ with $n > 2$ are assumed to be negligible and are set to zero. The system being of infinite order, an infinite number of harmonics is also generated for the reflected waves. The first five reflected waves can then be represented using the following Volterra series expansion:

$$b(0) = V_{0,1}(\mathbf{x}_2), \tag{6.9}$$

$$b(\omega_0) = a(\omega_0) \cdot V_{1,1}(\mathbf{x}_2) + a^*(\omega_0)a(2\omega_0) \cdot V_{1,2}(\mathbf{x}_2), \tag{6.10}$$

$$b(2\omega_0) = a(2\omega_0) \cdot V_{2,1}(\mathbf{x}_2) + a^2(\omega_0) \cdot V_{2,2}(\mathbf{x}_2), \tag{6.11}$$

$$\begin{aligned} b(3\omega_0) = {} & a^3(\omega_0) \cdot V_{3,1}(\mathbf{x}_2) + a(\omega_0)a(2\omega_0) \cdot V_{3,2}(\mathbf{x}_2) \\ & + a^*(\omega_0)a^2(2\omega_0) \cdot V_{3,3}(\mathbf{x}_2), \end{aligned} \tag{6.12}$$

$$\begin{aligned} b(4\omega_0) = {} & a^4(\omega_0) \cdot V_{4,1}(\mathbf{x}_2) + a^2(2\omega_0) \cdot V_{4,2}(\mathbf{x}_2) \\ & + a^2(\omega_0)a(2\omega_0) \cdot V_{4,3}(\mathbf{x}_2), \end{aligned} \tag{6.13}$$

$$\begin{aligned} b(5\omega_0) = {} & a^5(\omega_0) \cdot V_{5,1}(\mathbf{x}_2) + a(\omega_0)a^2(2\omega_0) \cdot V_{5,2}(\mathbf{x}_2) \\ & + a^*(\omega_0)a^3(2\omega_0) \cdot V_{5,3}(\mathbf{x}_2) + a^3(\omega_0)a(2\omega_0) \cdot V_{5,4}(\mathbf{x}_2). \end{aligned} \tag{6.14}$$

The functions $V_{n,i}(\mathbf{x}_2)$ in Equations (6.10)–(6.12) are found to be functionally dependent on five DC terms $x_{2,i}$:

$$V_{n,i}(\mathbf{x}_2) = V_{n,i}(x_{2,0}, x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}), \tag{6.15}$$

where these DC terms are given by

$$\begin{aligned} x_{2,0} &= a(0) = V_0, \\ x_{2,1} &= a^*(\omega_0)a(\omega_0), \\ x_{2,2} &= a^*(2\omega_0)a(2\omega_0), \\ x_{2,3} &= a^*(2\omega_0)a^2(\omega_0), \\ x_{2,4} &= x_{2,3}^* = a(2\omega_0)a^{*2}(\omega_0). \end{aligned}$$

Finally, let us consider injecting a signal consisting of three incident harmonic waves $a(\omega_0)$, $a(2\omega_0)$, and $a(3\omega_0)$ on the nonlinear impedance of a SISO circuit with infinite order. All other incident harmonics $a(n\omega_0)$ with $n > 3$ are assumed to be negligible and are set to zero. The first five reflected waves are represented using the following Volterra series expansion:

$$b(0) = V_{0,1}(\mathbf{x}_3), \tag{6.16}$$

$$\begin{aligned} b(\omega_0) = {} & a(\omega_0) \cdot V_{1,1}(\mathbf{x}_3) + a^*(\omega_0)a(2\omega_0) \cdot V_{1,2}(\mathbf{x}_3) \\ & + a^*(2\omega_0)a(3\omega_0) \cdot V_{1,3}(\mathbf{x}_3) + a^{*2}(\omega_0)a(3\omega_0) \cdot V_{1,4}(\mathbf{x}_3) \\ & + a^*(3\omega_0)a^2(2\omega_0) \cdot V_{1,5}(\mathbf{x}_3), \end{aligned} \tag{6.17}$$

$$\begin{aligned} b(2\omega_0) = {} & a^2(\omega_0) \cdot V_{2,1}(\mathbf{x}_3) + a(2\omega_0) \cdot V_{2,2}(\mathbf{x}_3) \\ & + a^*(2\omega_0)a(\omega_0)a(3\omega_0) \cdot V_{2,3}(\mathbf{x}_3) + a^*(\omega_0)a(3\omega_0) \cdot V_{2,4}(\mathbf{x}_3) \\ & + a^{*2}(2\omega_0)a^2(3\omega_0) \cdot V_{2,5}(\mathbf{x}_3), \end{aligned} \tag{6.18}$$

$$\begin{aligned} b(3\omega_0) = {} & a^3(\omega_0) \cdot V_{3,1}(\mathbf{x}_3) + a(\omega_0)a(2\omega_0) \cdot V_{3,2}(\mathbf{x}_3) \\ & + a^*(\omega_0)a^2(2\omega_0) \cdot V_{3,3}(\mathbf{x}_3) + a(3\omega_0) \cdot V_{3,4}(\mathbf{x}_3) \\ & + a^*(3\omega_0)a^3(2\omega_0) \cdot V_{3,5}(\mathbf{x}_3), \end{aligned} \tag{6.19}$$

$$\begin{aligned} b(4\omega_0) = {} & a^4(\omega_0) \cdot V_{4,1}(\mathbf{x}_3) + a^2(2\omega_0) \cdot V_{4,2}(\mathbf{x}_3) \\ & + a^2(\omega_0)a(2\omega_0) \cdot V_{4,3}(\mathbf{x}_3) + a(\omega_0)a(3\omega_0) \cdot V_{4,4}(\mathbf{x}_3) \\ & + a^*(2\omega_0)a^2(3\omega_0) \cdot V_{4,5}(\mathbf{x}_3) + a^*(\omega_0)a(2\omega_0)a(3\omega_0) \cdot V_{4,6}(\mathbf{x}_3) \\ & + a^{*2}(\omega_0)a^2(3\omega_0) \cdot V_{4,7}(\mathbf{x}_3), \end{aligned} \tag{6.20}$$

$$\begin{aligned} b(5\omega_0) = {} & a^5(\omega_0) \cdot V_{5,1}(\mathbf{x}_3) + a(\omega_0)a^2(2\omega_0) \cdot V_{5,2}(\mathbf{x}_3) \\ & + a^*(\omega_0)a^3(2\omega_0) \cdot V_{5,3}(\mathbf{x}_3) + a^3(\omega_0)a(2\omega_0) \cdot V_{5,4}(\mathbf{x}_3) \\ & + a(2\omega_0)a(3\omega_0) \cdot V_{5,5}(\mathbf{x}_3) + a^*(\omega_0)a^2(3\omega_0) \cdot V_{5,6}(\mathbf{x}_3) \\ & + a^2(\omega_0)a(3\omega_0) \cdot V_{5,7}(\mathbf{x}_3) + a^*(2\omega_0)a(\omega_0)a^2(3\omega_0) \cdot V_{5,8}(\mathbf{x}_3) \\ & + a^*(3\omega_0)a^4(2\omega_0) \cdot V_{5,9}(\mathbf{x}_3) + a^{*2}(2\omega_0)a^3(3\omega_0) \cdot V_{5,10}(\mathbf{x}_3). \end{aligned} \tag{6.21}$$

The functions $V_{p,i}(\mathbf{x}_3)$ in Equations (6.16)–(6.21) for the first five harmonics and all other harmonics p are found to be functionally dependent on 14 DC terms $x_{3,i}$:

$$V_{p,i}(\mathbf{x}_3) = V_{p,i}(x_{3,0}, x_{3,1}, x_{3,2}, \ldots, x_{3,13}), \tag{6.22}$$

where these DC terms are given by

$$\begin{aligned} & x_{3,0} = a(0) = V_0, \\ & x_{3,1} = a^*(\omega_0)a(\omega_0), && x_{3,2} = a^*(2\omega_0)a(2\omega_0), \\ & x_{3,3} = a^*(2\omega_0)a^2(\omega_0), && x_{3,4} = x_{3,3}^*, \\ & x_{3,5} = a^*(3\omega_0)a(3\omega_0), \\ & x_{3,6} = a^*(3\omega_0)a(\omega_0)a(2\omega_0), && x_{3,7} = x_{3,6}^*, \\ & x_{3,8} = a^*(3\omega_0)a^3(\omega_0), && x_{3,9} = x_{3,8}^*, \\ & x_{3,10} = a^*(3\omega_0)a^*(\omega_0)a^2(2\omega_0), && x_{3,11} = x_{3,10}^*, \\ & x_{3,12} = a^{*2}(3\omega_0)a^3(2\omega_0), && x_{3,13} = x_{3,12}^*. \end{aligned}$$

This expansion for three harmonic excitations holds up to infinite order in nonlinearity but obviously requires that the $V_{p,i}(\mathbf{x}_3)$ functions be identified.

The expansion for four harmonic excitations is presented in Section 6.5. Generally speaking the harmonic $p\omega_0$ for n incident harmonic tones can be expressed in terms of a finite series expansion of length $M_{n,p}$:

$$
\begin{aligned}
b(0) &= V_{0,1}(\mathbf{x}_n), \\
b(p\omega_0) &= \sum_{m=1}^{M_{n,p}} x_{n,m}(p) V_{p,m}(\mathbf{x}_n) \qquad \text{for} \quad p > 0.
\end{aligned}
\tag{6.23}
$$

In the next section we discuss the method used to derive the above compact results for a system with infinite order excited by a finite number n of harmonic excitations.

6.2.2 Model derivation

All the fundamental and DC terms in Equations (6.3) and (6.21) from one up to three harmonics and in Equations (6.68) and (6.70) for four harmonics in Section 6.5 can be obtained from an exhaustive search using a simple MATLAB algorithm. Let us first briefly explain the guiding principle for the algorithm used to find the DC terms in Equations (6.15), (6.22), and (6.72).

For an infinite-order nonlinear system excited by n incident harmonic tones, a matrix $\mathbf{D}_{n,p}$ for the p^{th} harmonic can be introduced to record the power (exponent) $D_{n,p}(n+1+k,i)$ of each tone $a(k\omega_0)$ contributing to term i of the p^{th}-harmonic term. It results that we have, using $a(-k\omega_0) = a^*(k\omega_0)$,

$$
x_{n,i}(p) = \prod_{k=-n}^{n} [a(k\omega_0)]^{D_{n,p}(n+1+k,i)},
$$

where we define $x_{n,m}(0) = x_{n,m}$ for the DC terms ($p = 0$). The frequency of each tone k is given by the element $F_n(k)$ of the companion vector $\mathbf{F}_n$:

$$
\mathbf{F}_n = \begin{bmatrix} -n & \dots & -k & \dots & -2 & -1 & 0 & 1 & 2 & \dots & k & \dots & n \end{bmatrix}.
$$

The matrix $\mathbf{D}_{n,p}$ for the p^{th} harmonic satisfies therefore

$$
\mathbf{F}_n \times \mathbf{D}_{n,p} = \begin{bmatrix} p & \dots & p & p & p & p & p & \dots & p \end{bmatrix} = \mathbf{p}_{M_{n,p}}.
$$

The total order for a term $x_{n,i}(p)$ is then given by the sum of the exponent of its components:

$$
O_{n,p}(i) = \sum_{k=1}^{2n+1} D_{n,p}(k,i).
$$

For example, for $n = 3$ controlled harmonics according to the definitions given in Equation (6.21), the following 7×14 exponent matrix $\mathbf{D}_{3,0}$ is obtained to define the DC terms:

$$
\mathbf{D}_{3,0} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 2 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 3 \\
0 & 1 & 0 & 0 & 2 & 0 & 0 & 1 & 0 & 3 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 2 & 0 & 0 & 1 & 0 & 3 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 3 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 2
\end{bmatrix}.
\tag{6.24}
$$

In the exponent matrix $\mathbf{D}_{3,0}$, each column represents a DC term and each row represents a harmonic frequency from $-3\omega_0$ (top) to $+3\omega_0$ (bottom), with the center row associated with the DC component (frequency $0 \times \omega_0$). Note that all the coefficients in the exponent matrix $\mathbf{D}_{3,0}$ are positive or null integers. Negative integers would be conceptually associated with dividers, but sub-harmonic generation is outside the class of systems considered here.

Let $\mathbf{F}_3$ be the harmonic frequency vector:

$$\mathbf{F}_3 = \begin{bmatrix} -3 & -2 & -1 & 0 & 1 & 2 & 3 \end{bmatrix}.$$

The $x_{3,i}$ columns retained in the exponent matrix $\mathbf{D}_{3,0}$ are those which generate DC contributions by verifying the condition

$$\mathbf{F}_3 \times \mathbf{D}_{3,0} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \mathbf{0}_{14}$$

and which cannot be generated from the superposition of other lower-order $x_{3,i}$ exponent columns. Indeed, infinitely many other $x_{3,i}$ featuring higher orders are possible but can be factorized in terms of the previously extracted lower-order DC terms $x_{3,i}$. For example, consider the following randomly generated high-order DC term:

$$y(0) = a(3\omega)^{*77} \cdot a(2\omega)^{*80} \cdot a(\omega)^{*19} \cdot a(0\omega)^{0} \cdot a(\omega)^{45} \cdot a(2\omega)^{76} \cdot a(3\omega)^{71},$$

represented by its power-exponent vector $\mathbf{Y}_0$,

$$\mathbf{Y}_0 = \begin{bmatrix} 77 & 80 & 19 & 0 & 45 & 76 & 71 \end{bmatrix}^{\mathsf{T}}.$$

Since $\mathbf{Y}_0$ is a DC term, it satisfies

$$\mathbf{F}_3 \times \mathbf{Y}_0 = 0.$$

$\mathbf{Y}_0$ can be uniquely expressed in the $\mathbf{D}_{3,0}$ basis,

$$\mathbf{Y}_0 = \mathbf{D}_{3,0} \times \mathbf{X}_0,$$

using the following $\mathbf{X}_0$:

$$\mathbf{X}_0 = \begin{bmatrix} 0 & 19 & 76 & 4 & 0 & 71 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^{\mathsf{T}}.$$

It results that the higher-order DC term y can be rewritten in the $\mathbf{x}_3$ basis as

$$y(0) = x_{3,1}^{19} \cdot x_{3,2}^{76} \cdot x_{3,3}^{4} \cdot x_{3,5}^{71} \cdot x_{3,8}^{6},$$

and is therefore not a new independent DC term.

Using a similar methodology, an exponent matrix $\mathbf{D}_{3,1}$ is used to represent the exponent (power) of the five fundamental terms $x_{3,1}(1) = a(\omega_0)$, $x_{3,2}(1) = a^*(\omega_0)a(2\omega_0)$, $x_{3,3}(1) = a^*(2\omega_0)a(3\omega_0)$, $x_{3,4}(1) = a^{*2}(\omega_0)a(3\omega_0)$, and $x_{3,5}(1) = a^*(3\omega_0)a^2(2\omega_0)$ involved in the definition of $b(\omega_0)$ in Equation (6.17) for the fundamental frequency $1 \times \omega_0$:

$$\mathbf{D}_{3,1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}. \tag{6.25}$$

The $y_{3,i}$ columns retained in the exponent matrix $\mathbf{D}_{3,1}$ are those which generate fundamental frequency contributions by satisfying the condition

$$\mathbf{F}_3 \times \mathbf{D}_{3,1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \mathbf{1}_5,$$

and which cannot be further reduced by the factorization of DC terms $x_{3,i}$ from the exponent matrix $\mathbf{D}_{3,0}$.

For example, consider the following randomly generated high-order fundamental term (ω_0):

$$y(\omega_0) = a(3\omega_0)^{*81} \cdot a(2\omega_0)^{*91} \cdot a(\omega_0)^{*13} \cdot a(0\omega_0)^{0} \cdot a(\omega_0)^{63} \cdot a(2\omega_0)^{146} \cdot a(3\omega_0)^{28}.$$

Its exponent vector $\mathbf{Y}_1$ satisfies

$$\mathbf{F}_3 \times \mathbf{Y}_1 = 1.$$

The exponent vector $\mathbf{Y}_1$ can then be uniquely expanded in the $\mathbf{D}_{3,0}$ and $\mathbf{D}_{3,1}$ bases using

$$\mathbf{Y}_1 = \begin{bmatrix} 81 \\ 91 \\ 13 \\ 0 \\ 63 \\ 146 \\ 28 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 2 \\ 0 \end{bmatrix} + \mathbf{D}_{3,0} \times \mathbf{X}_1,$$

using the following vector $\mathbf{X}_1$:

$$\mathbf{X}_1 = \begin{bmatrix} 0 & 13 & 91 & 0 & 0 & 28 & 50 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}^{\mathsf{T}}.$$

It results that the higher-order DC term $y(\omega_0)$ can be rewritten in terms of $x_{3,5}(1)$ (the fifth column of $\mathbf{D}_{3,1}$) and the $\mathbf{x}_3$ DC basis ($\mathbf{D}_{3,0}$) as

$$y(\omega_0) = x_{3,5}(1) \times \left[x_{3,1}^{13}(0) \cdot x_{3,2}^{91}(0) \cdot x_{3,5}^{28}(0) \cdot x_{3,6}^{50}(0) \cdot x_{3,12}^{1}(0) \right],$$

and is therefore not a new independent fundamental (ω_0) term.

The methodology given above for the case of three tones can be implemented for a system with an arbitrary number n of controlled incident harmonic tones. The MATLAB script used to obtain the DC, fundamental, and higher-harmonic coefficients will be posted on the webpage for this book. These results indicate that for a finite number n of

controlled harmonic excitations applied to an infinite-order nonlinear system exhibiting arbitrary nonlinearities of the form $y(p\omega_0)$ defined by

$$y(p) = \prod_{k=-n}^{n} [a(k\omega_0)]^{Y_p(k)},$$

with power exponents given by a $(2n+1) \times 1$ vector $\mathbf{Y}_p$ for the p^{th} harmonic, thus satisfying the frequency constraint $\mathbf{F}_n \times \mathbf{Y}_p = \mathbf{p}$, there exists a vector $\mathbf{X}_p$ of positive integers and a column index i such that we have

$$\mathbf{Y}_p = \mathbf{D}_{n,p}(:, i) + \mathbf{D}_{n,0} \times \mathbf{X}_p,$$

where the finite $(2n+1) \times M_{n,0}$ exponent matrix $\mathbf{D}_{n,0}$ and $(2n+1) \times M_{n,p}$ exponent matrix $\mathbf{D}_{n,p}$ are finite and invariant bases for the DC and p harmonic terms, respectively. Using this asserted number-theory result, the validity of the finite expansion postulated in Equation (6.23) for an infinite-order system can then be established to the extent that the Volterra functions $V_{p,m}(\mathbf{x}_n)$ thus defined admit a converging Maclaurin expansion (entire functions) relative to the complex variables $\mathbf{x}_n$.

6.2.3 Analytic example

As indicated above, the model derived holds up to infinite order, assuming that the nonlinearity can be represented by an infinite power series. As an example of a SISO system with infinite-order nonlinearity, let us consider an ideal diode:

$$i_{\text{D}}(v_{\text{D}}) = I_0\big[\exp(\alpha v_{\text{D}}) - 1\big],$$

with $\alpha = 1/(kT)$. We assume for the sake of simplicity that the applied voltage consists of the fundamental tone plus the second harmonic (case $n = 2$):

$$\begin{aligned} v_{\text{D}}(t) &= V_0 + V_1 \cos(\omega_0 t + \phi_1) + V_2 \cos(2\omega_0 t + \phi_2) \\ &= V_0 + 2\,\text{Re}\,\{a(\omega_0)\exp(j\omega_0 t) + a(2\omega_0)\exp(j2\omega_0 t)\}, \end{aligned}$$

using

$$a(\omega_0) = \frac{V_1}{2}\exp(j\phi_1) \qquad \text{and} \qquad a(2\omega_0) = \frac{V_2}{2}\exp(j\phi_2).$$

The diode current is then

$$i_{\text{D}}(t) = \sum_{k=-\infty}^{\infty} b(k\omega_0)\exp(jk\omega_0 t).$$

Using the relation

$$\exp\big[\alpha A_p \cos(p\omega_0 t + \phi_p)\big] = \sum_{n=-\infty}^{\infty} I_n[\alpha A_p]\exp\big[jn(p\omega_0 t + \phi_p)\big],$$

where $I_n(x)$ are the modified Bessel functions of the first kind,

$$I_n[x] = I_{-n}[x] = \left(\frac{x}{2}\right)^{|n|} \sum_{p=0}^{\infty} \frac{(x/2)^{2p}}{p!(|n|+p)!}$$
$$= (x)^{|n|}\, \overline{I}_n \left[\left(\frac{x}{2}\right)^2\right], \tag{6.26}$$

the functions $V_{n,i}(\mathbf{x}_2)$ introduced in Equations (6.10)–(6.12) for the first three harmonics can now be derived to be

$$V_{0,1}(\mathbf{x}_2) = C_0 \sum_{n=-\infty}^{-1} (2\alpha)^{|3n|} x_{3,2}^{|n|}\, \overline{I}_{-2n}[\alpha^2 x_{1,2}]\, \overline{I}_n[\alpha^2 x_{2,2}]$$
$$+ C_0 \sum_{n=0}^{\infty} (2\alpha)^{|3n|} x_{4,2}^{n}\, \overline{I}_{-2n}[\alpha^2 x_{1,2}]\, \overline{I}_n[\alpha^2 x_{2,2}],$$
$$V_{1,1}(\mathbf{x}_2) = C_0 \sum_{n=-\infty}^{0} (2\alpha)^{|1-3n|} x_{3,2}^{|n|} \overline{I}_{1-2n}[\alpha^2 x_{1,2}]\, \overline{I}_n[\alpha^2 x_{2,2}],$$
$$V_{1,2}(\mathbf{x}_2) = C_0 \sum_{n=1}^{\infty} (2\alpha)^{|1-3n|} x_{4,2}^{n-1} \overline{I}_{1-2n}[\alpha^2 x_{1,2}] \overline{I}_n[\alpha^2 x_{2,2}],$$
$$V_{2,1}(\mathbf{x}_2) = C_0 \sum_{n=1}^{\infty} (2\alpha)^{|2-3n|} x_{4,2}^{n-1} \overline{I}_{2-2n}[\alpha^2 x_{1,2}] \overline{I}_n[\alpha^2 x_{2,2}],$$
$$V_{2,2}(\mathbf{x}_2) = C_0 \sum_{n=-\infty}^{0} (2\alpha)^{|2-3n|} x_{3,2}^{|n|}\, \overline{I}_{2-2n}[\alpha^2 x_{1,2}]\, \overline{I}_n[\alpha^2 x_{2,2}],$$
$$V_{3,1}(\mathbf{x}_2) = C_0 \sum_{n=-\infty}^{0} (2\alpha)^{|3-3n|} x_{3,2}^{|n|} \overline{I}_{3-2n}[\alpha^2 x_{1,2}] \overline{I}_n[\alpha^2 x_{2,2}],$$
$$V_{3,2}(\mathbf{x}_2) = C_0\ (2\alpha)^2 \overline{I}_1[\alpha^2 x_{1,2}] \overline{I}_1[\alpha^2 x_{2,2}],$$
$$V_{3,3}(\mathbf{x}_2) = C_0 \sum_{n=2}^{\infty} (2\alpha)^{|3-3n|} x_{4,2}^{n-2}\, \overline{I}_{3-2n}[\alpha^2 x_{1,2}]\, \overline{I}_n[\alpha^2 x_{2,2}],$$

with $C_0 = I_0 \exp(\alpha V_0)$. Note that in practice an infinite summation over n is not needed since convergence occurs rapidly. A summation for n from -10 to 10 is sufficient for the range of the diode voltage used (0.5 V). Figure 6.2 shows the Volterra function $V_{n,m}(\mathrm{x}_2)$ parametrically plotted versus the amplitude V_1. The Volterra functions are complex nonlinear functions of their variables x_p.

Figure 6.3 shows the error for the fundamental harmonic versus V_2/V_1 for various orders of the Volterra expansion for the second harmonics while using an infinite order for the fundamental. For a V_2/V_1 ratio larger than 20 dB, the error becomes large for this exponentional nonlinearity. The Volterra function expansion (squares) yields no distinguishable error.

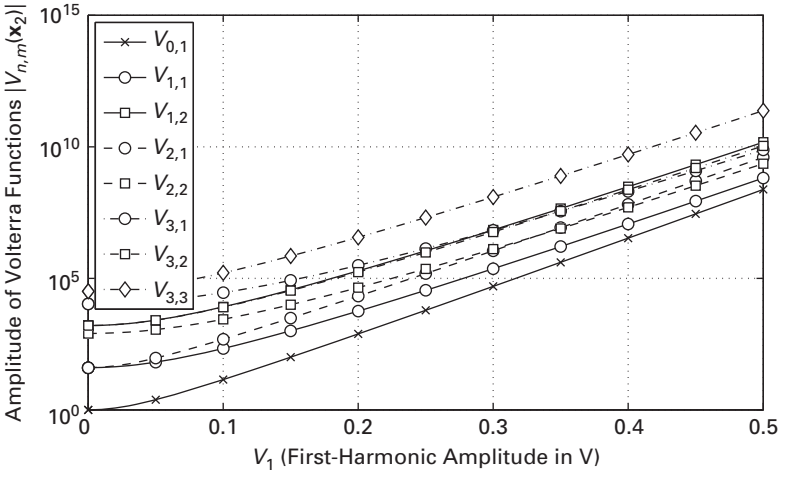


Figure 6.2 The amplitudes of the Volterra functions $|V_{n,m}(\mathbf{x}_2)|$ versus the parametric variable V_1 for $V_2 = V_1/10$ and $\phi_2 = \pi/7$.

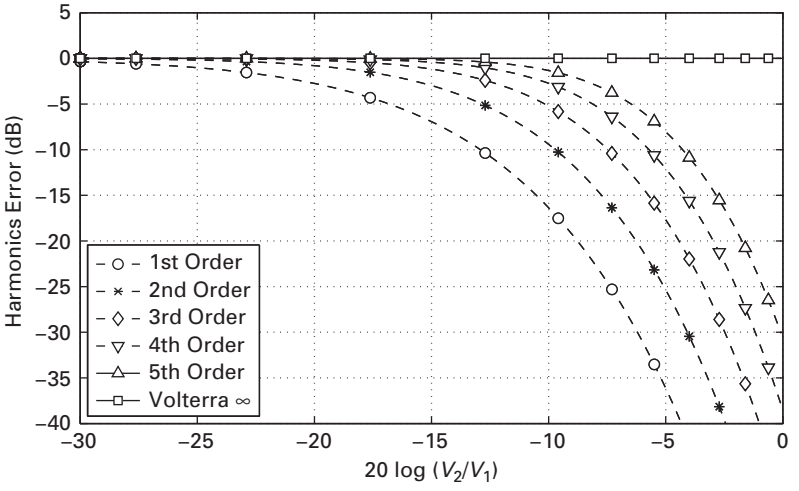


Figure 6.3 Error for the fundamental harmonic versus V_2/V_1 for a DC bias of $V_0 = 0.3$ V. The fundamental amplitude $V_1 = 0.3$ V and $\phi_2 = \pi/7$. It can be verified that a similar error is obtained for the first five harmonics (not shown).

Figure 6.4 shows the waveforms obtained for driving voltages V_1 of 0.03 V and 0.3 V while keeping the second-to-first-harmonic ratio at -10 dBc. A first-order power series of the second harmonic is seen to be acceptable under the small-signal excitation 0.03 V but not for the large-signal excitation 0.3 V. The Volterra function expansion (squares) is in agreement with the exact waveform.

The model presented above is memoryless. A simple model with memory can be constructed simply by adding in shunt to the diode current the displacement current obtained from the storage capacitance of the ideal diode:

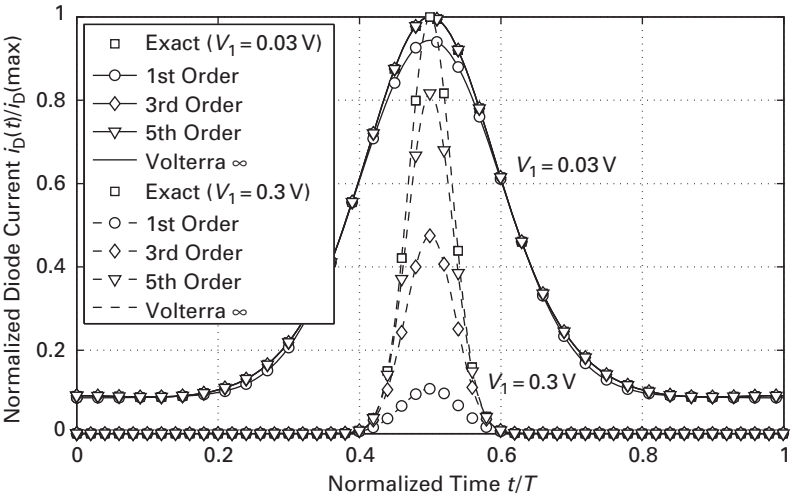


Figure 6.4 Current waveform normalized with respect to its peak for a DC bias $V_0 = 0.3$ V and fundamental amplitudes V_1 of 0.03 V and 0.3 V. The harmonic ratio V_2/V_1 is kept equal to $0.316 = -10$ dBc.

$$i_Q = \frac{dQ_D(v)}{dt} = \frac{1}{\tau}\frac{di_D(v_D)}{dt} = \frac{1}{\tau}\frac{di_D(v_D)}{dv_D}\frac{dv_D}{dt}.$$

The formula derived above for i_D holds now for the charge $Q_D = i_D/\tau$. Similar results for the displacement current are therefore readily demonstrated with this model exhibiting memory with infinite-order nonlinearity.

6.2.4 Model extraction

The theory developed in the previous section holds up to infinite order for a finite number of harmonics n but requires the identification of the Volterra $V_{p,i}(\mathbf{x}_n)$. This identification can be performed conceptually using modern nonlinear vector network analyzers, which acquire the phase and amplitude of the fundamental and harmonics of the incident and reflected waves and can therefore map the dependence of the reflected waves on the amplitude and phase of the incident waves. However, as indicated by the theory, as the number of incident harmonics p increases both the number of Volterra functions $V_{p,i}(\mathbf{x}_n)$ and the size of $\mathbf{x}_n$ increase. The exhaustive identification of the resulting multidimensional nonlinear response then quickly becomes laborious for a large number of incident harmonics.

For the case of mild nonlinearities it is possible to expand the Volterra $V_{p,i}(\mathbf{x}_n)$ in a truncated Maclaurin series around the DC operating point $[x_{n,0}]$:

$$\begin{aligned} V_{p,i}(\mathbf{x}_n) &= V_{p,i}(x_{n,0}, \ldots, x_{n,P}) = V_{p,i}(x_{n,0}, 0 \ldots, 0) \\ &+ \sum_{r_1=0}^{O_1} \cdots \sum_{r_P=0}^{O_P} \frac{x_{n,1}^{r_1} \ldots x_{n,P}^{r_P}}{n_1! \ldots n_P!} \times \frac{\partial^{r_1+\cdots+r_P} V_{n,i}}{\partial x_{n,1}^{r_1} \ldots \partial x_{n,P}^{r_P}}(x_{n,0}, 0 \ldots, 0). \quad (6.27) \end{aligned}$$

In the case of a mild nonlinear regime, higher-order nonlinearities can then be accounted for by properly selecting the expansion order $[O_1, \ldots, O_P]$ and pruning out the terms which are not contributing substantially to the power series. An example is given in the next section.

6.2.5 Experimental model extraction and validation

In this section we shall apply the extraction methodology presented in the previous section to the modeling of a stable nonlinear negative resistance (a one-port device). A specific target is the extraction of the Volterra model parameters needed to accurately represent the dependence of the fundamental power $\frac{1}{2}(|b(\omega_0)|^2 - |a(\omega_0)|^2)$ on the second-harmonic impedance termination acquired in a real-time active load-pull (RTALP) experiment.

We will first apply the two-harmonic analysis since only incident signals at the fundamental and second harmonic are applied, and the RF generators are expected to provide a reasonable match impedance termination at the higher frequencies. Both the fundamental $b(\omega_0)$ and the second-harmonic wave $b(2\omega_0)$ are used for the model extraction along the contour lines of the RTALP experiment shown in Figure 6.5. A good prediction of $b(2\omega_0)$ is obtained.

A contour plot of the device output power at the fundamental frequency as a function of the second-harmonic impedance is shown in Figure 6.6. The minimum power located at the cross is -16.2 dBm and the maximum power located at the plus is -13.7 dBm.

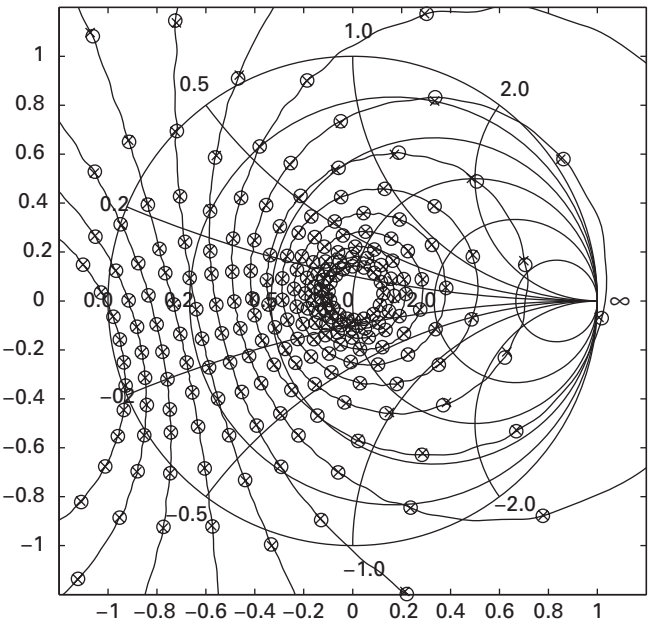


Figure 6.5 Comparison of the measured and extracted (crosses) loci of the $b(2\omega_0)/a(2\omega_0)$ second-harmonic reflection coefficient as a function of the incident waves $a(\omega_0)$ and $a(2\omega_0)$ in the RTALP experiment.

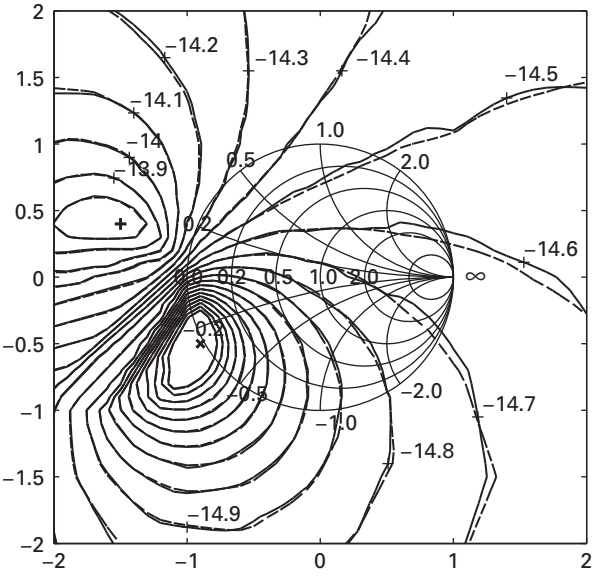


Figure 6.6 Contour plot with 0.1-dB steps of the device output power at the fundamental frequency as a function of the second-harmonic impedance. Measured and simulated data are represented by plain and dashed lines, respectively.

The tuning of the second harmonic is seen to provide a 2.5-dB increase in output power for the negative resistance. The model is found to be able to accurately render the RTALP contour plot even though the model was extracted over a relatively sparse mapping of the Smith chart given by 15 RTALP locus lines in Figure 6.5.

A high-order Volterra expansion was needed in order to obtain the results reported. The fundamental harmonic Volterra functions were expanded up to order 12 and the second-harmonic Volterra functions were expanded up to order 9. The device is therefore exhibiting a substantial nonlinearity in the RTALP testbed used when it is analyzed using two-harmonic tones.

To further test the model capability in predicting the device behavior, various power contour plots (dashed lines) are generated when a single locus line is skipped at a time in the measurement database shown in Figure 6.5. The resulting contour plots are superposed in Figure 6.7 and seen to depart only slighly from the one (plain lines) calculated directly from the measured data using the full database (all loci in Figure 6.5). Also, consistant values are verified to be extracted for most of the dominant model parameters.

It is equally important to note that, in the RTALP measurements, the effective impedance of the testbed generator at the fundamental frequency departs from an ideal match. The generator impedance is simply extracted from the RTALP data using

$$a(\omega_0, P, t) = \Gamma_G(\omega_0) \cdot b(\omega_0, P, t) + b_G(\omega_0, P),$$

where t is the RTALP sweep-time parameter. The amplitude of the generator reflection coefficient is found to be $|\Gamma_G(\omega_0)| = 0.081$. Having a non-zero $|\Gamma_G(\omega_0)|$ is quite significant because the modulation of $b(\omega_0, t)$ is reflected back into the device measured. Note that, as the second-harmonic incident wave $a(2\omega_0)$ is swept in phase versus

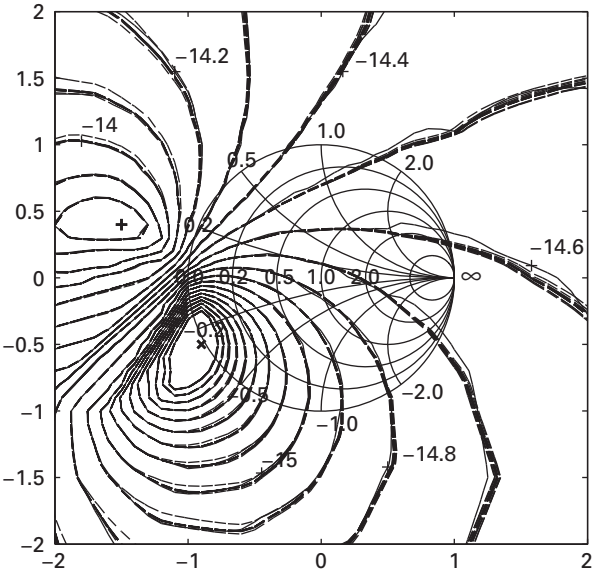


Figure 6.7 Verification of the stability of the extraction.

time t and the power P is stepped to different values, the reflected wave $b(\omega_0)$ at the fundamental is modulated owing, among other things, to the Volterra function $V_{1,2}(\mathbf{x}_2)$. The non-zero reflection coefficent $|\Gamma_G(\omega_0)|$ thus contributes an external feedback that increases the apparent nonlinearity of the device compared with the situation of an ideally matched generator. Because $a(\omega_0)$ is not held constant but $b_G(\omega_0, P)$ is for a given preselected input power P, we have elected to use the latter as the input variable in the above Volterra modeling. In this way somewhat smoother power contours result for the measured data shown in Figure 6.5.

To use this model for circuit simulations, the Volterra model has been implemented[2] in the Agilent Advanced Design System (ADS) using the built-in two-port frequency-domain-defined device (FDD) component. Owing to the automated MATLAB extraction developed, all the Volterra coefficients and associated theoretical expressions required for the nonlinear mixing terms are easily embedded into the simulator. The number of nonlinear coefficients was reduced in the ADS simulations to prevent the occurrence of any convergence problems.

6.2.6 Phase reference

As was discussed in Chapter 2, the nonlinear frequency-domain data are acquired by the NVNA up to an arbitrary phase due to the arbitrary time reference t' of the measurement:

[2] Unpublished work with Inwon Suh.

$$a(t') = \mathrm{Re}\left\{\sum_p a'(p\omega_0)\exp\left[j\left(p\omega_0 t' + \phi'_p\right)\right]\right\}.$$

For example, for a SIMO system, consider the various harmonics measured with their time dependence:

$$\begin{aligned}
a'(\omega_0)\exp\left(j\omega_0 t'\right) &= |a'(\omega_0)|\exp\left[j\left(\omega_0 t' + \phi'_1\right)\right],\\
a'(2\omega_0)\exp\left(j2\omega_0 t'\right) &= |a'(2\omega_0)|\exp\left[j\left(2\omega_0 t' + \phi'_2\right)\right],\\
a'(3\omega_0)\exp\left(j3\omega_0 t'\right) &= |a'(3\omega_0)|\exp\left[j\left(3\omega_0 t' + \phi'_3\right)\right],\\
&\cdots\\
a'(n\omega_0)\exp\left(jn\omega_0 t'\right) &= |a'(n\omega_0)|\exp\left[j\left(n\omega_0 t' + \phi'_n\right)\right].
\end{aligned}$$

By moving the time origin from t' to $t = t' + \tau$ using the time shift $\tau = \phi_1/\omega_0$ such that we have $\omega_0(t-\tau) + \phi_1 = \omega_0 t$ we obtain the following phasors:

$$\begin{aligned}
a(\omega_0) &= |a(\omega_0)|,\\
a(2\omega_0) &= |a(2\omega_0)|\exp(j\phi_2),\\
a(3\omega_0) &= |a(3\omega_0)|\exp(j\phi_3),\\
&\cdots\\
a(n\omega_0) &= |a(n\omega_0)|\exp(j\phi_n),
\end{aligned}$$

with the same amplitude $|a(n\omega_0)| = |a'(n\omega_0)|$ and the new phases:

$$\begin{aligned}
\phi_1 &= 0,\\
\phi_2 &= \phi'_2 - 2\omega_0\tau = \phi'_2 - 2\phi'_1,\\
\phi_3 &= \phi'_3 - 3\omega_0\tau = \phi'_3 - 3\phi'_1,\\
&\cdots\\
\phi_n &= \phi'_n - n\omega_0\tau = \phi'_n - n\phi'_1.
\end{aligned} \tag{6.28}$$

We shall assume in the rest of this work that such a trivial operation was performed on the measured data such that we now have, as needed,

$$a(\omega) = |a(\omega)|$$

at the excitation port of the SIMO system. In a MIMO situation the above identity can be verified only at a single input port r selected to be port 1 for simplicity so that we have $a_r(\omega) = |a_r(\omega)|$ only for $r = 1$.

6.2.7 Poly-harmonic distortion model (PHD)

The poly-harmonic distortion (PHD) [9] approximation retains only the harmonic ($p > 1$) terms $a(p\omega_0)$ and $a^*(p\omega_0)$ with a first-order power while using a high-order (as large as needed) expansion for the fundamental frequency. This approximation is readily applied to the exact Volterra function derived above. We first present the two-harmonic excitation derivation:

$$b(0) = \left[V_{0,1}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{0,1}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{0,1}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right], \tag{6.29}$$

$$\begin{aligned} b(\omega_0) = {} & a(\omega_0) \cdot \left[V_{1,1}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{1,1}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{1,1}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right] \\ & + a^*(\omega_0) a(2\omega_0) \cdot V_{1,2}\left(\mathbf{x}_2^{(0)}\right), \end{aligned} \tag{6.30}$$

$$\begin{aligned} b(2\omega_0) = {} & a(2\omega_0) \cdot V_{2,1}(\mathbf{x}_2^{(0)}) \\ & + a^2(\omega_0) \left[V_{2,2}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{2,2}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{2,2}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right], \end{aligned} \tag{6.31}$$

$$\begin{aligned} b(3\omega_0) = {} & a^3(\omega_0) \cdot \left[V_{3,1}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{3,1}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{3,1}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right] \\ & + a(\omega_0) a(2\omega_0) \cdot V_{3,2}\left(\mathbf{x}_2^{(0)}\right), \end{aligned} \tag{6.32}$$

$$\begin{aligned} b(4\omega_0) = {} & a^4(\omega_0) \cdot \left[V_{4,1}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{4,1}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{4,1}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right] \\ & + a^2(\omega_0) \cdot \left[V_{4,2}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{4,2}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{4,2}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right] \\ & + a^2(\omega_0) a(2\omega_0) \cdot V_{4,3}\left(\mathbf{x}_2^{(0)}\right), \end{aligned} \tag{6.33}$$

$$\begin{aligned} b(5\omega_0) = {} & a^5(\omega_0) \left[V_{5,1}\left(\mathbf{x}_2^{(0)}\right) + \frac{\partial V_{5,1}}{\partial x_{2,3}}\left(\mathbf{x}_2^{(0)}\right) x_{2,3} + \frac{\partial V_{5,1}}{\partial x_{2,4}}\left(\mathbf{x}_2^{(0)}\right) x_{2,4} \right] \\ & + a^3(\omega_0) a(2\omega_0) \cdot V_{5,4}\left(\mathbf{x}_2^{(0)}\right), \end{aligned} \tag{6.34}$$

where we have defined the new following operating point $\mathbf{x}_2^{(0)}$ to be

$$\mathbf{x}_2^{(0)} = (x_{2,0}, x_{2,1}, 0, 0, 0).$$

Note that the only DC terms needed in this expansion are

$$\begin{aligned} x_{2,0} &= V_0, & x_{2,1} &= a^*(\omega_0) a(\omega_0), \\ x_{2,3} &= a^*(2\omega_0) a^2(\omega_0) & x_{2,4} &= a(2\omega_0) a^{*2}(\omega_0). \end{aligned}$$

The expansion derived above for two-harmonic excitation can easily be extended to more harmonics. For example, the reflected wave $b(\omega_0)$ for four incident harmonics can be represented using the following Volterra series expansion:

$$\begin{aligned} b(\omega_0) = {} & a(\omega_0) \cdot V_{1,1}\left(\mathbf{x}_4^{(0)}\right) \\ & + a(\omega_0) \cdot \left[\frac{\partial V_{1,1}}{\partial x_{4,3}}\left(\mathbf{x}_4^{(0)}\right) x_{4,3} + \frac{\partial V_{1,1}}{\partial x_{4,8}}\left(\mathbf{x}_4^{(0)}\right) x_{4,8} + \frac{\partial V_{1,1}}{\partial x_{4,23}}\left(\mathbf{x}_4^{(0)}\right) x_{4,23} \right] \\ & + a(\omega_0) \cdot \left[\frac{\partial V_{1,1}}{\partial x_{4,4}}\left(\mathbf{x}_4^{(0)}\right) x_{4,4} + \frac{\partial V_{1,1}}{\partial x_{4,9}}\left(\mathbf{x}_4^{(0)}\right) x_{4,9} + \frac{\partial V_{1,1}}{\partial x_{4,24}}\left(\mathbf{x}_4^{(0)}\right) x_{4,24} \right] \\ & + a^*(\omega_0) a(2\omega_0) \cdot V_{1,2}\left(\mathbf{x}_4^{(0)}\right) + a^{*2}(\omega_0) a(3\omega_0) \cdot V_{1,4}\left(\mathbf{x}_4^{(0)}\right) \\ & + a^{*3}(\omega_0) a(4\omega_0) \cdot V_{1,8}\left(\mathbf{x}_4^{(0)}\right), \end{aligned} \tag{6.35}$$

where the operating point $\left(\mathbf{x}_4^{(0)}\right)$ for this perturbative analysis is

$$\mathbf{x}_4^{(0)} = (x_{4,0}, x_{4,1}, 0, 0, 0) = (V_0, |a(\omega_0)|^2, 0, \dots).$$

Note that, of the original 34 DC terms x_4, the only ones contributing in this first-order pertubative analysis are the following 7 DC terms $x_{4,i}$:

$$\begin{aligned}
&x_{4,1} = a^*(\omega_0)a(\omega_0),\\
&x_{4,3} = a^*(2\omega_0)a^2(\omega_0), \quad x_{4,4} = a(2\omega_0)a^{*2}(\omega_0),\\
&x_{4,8} = a^*(3\omega_0)a^3(\omega_0), \quad x_{4,9} = a(3\omega_0)a^{*3}(\omega_0),\\
&x_{4,23} = a^*(4\omega_0)a^4(\omega_0), \quad x_{4,24} = a(4\omega_0)a^{*4}(\omega_0).
\end{aligned}$$

On generalizing these results to an arbitrary number n of harmonic excitations we can recast this pertubative result in the traditional PHD form [9]:

$$b_r(p\omega_0) = \sum_{p=1}^{n} S_{r,1;p,q}(V_0, |a_1(\omega_0)|)a_1(q\omega_0) + T_{r,1;p,q}(V_0, |a_1(\omega_0)|)a_1^*(q\omega_0).$$

Note that no phasor appears in this derivation as in [9] since we have taken the trivial step described in the previous section of changing the time reference so that we have $a_1(\omega_0) = |a_1(\omega_0)|$ at port 1. The indices 1 and r were also added to indicate the input port 1 and the output port r, respectively, for the SIMO system considered. For a MIMO system with S input ports, this equation is generalized to

$$b_r(p\omega_0) = \sum_{s=1}^{S}\sum_{p=1}^{n} S_{r,s;p,q}(V_0, |a_1(\omega_0)|)a_s(q\omega_0) + T_{r,s;p,q}(V_0, |a_1(\omega_0)|)a_s^*(q\omega_0).$$

For the four-tone SISO system analyzed perturbatively we have derived the following results for the PHD parameters: $S_{r,1;p,q}(V_0, |a_1(\omega_0)|)$ and $T_{r,1;p,q}(V_0, |a_1(\omega_0)|)$, where we have

$$\begin{aligned}
&S_{r,1;1,1}(V_0, |a_1(\omega_0)|) = a_1(\omega_0) \cdot V_{1,1}\left(\mathbf{x}_4^{(0)}\right),\\
&T_{r,1;1,1}(V_0, |a_1(\omega_0)|) = 0,\\
&S_{r,1;1,2}(V_0, |a_1(\omega_0)|) = a_1(\omega_0)a_1^{*2}(\omega_0) \cdot \frac{\partial V_{1,1}}{\partial x_{4,4}}\left(\mathbf{x}_4^{(0)}\right) + a_1^*(\omega_0) \cdot V_{1,2}\left(\mathbf{x}_4^{(0)}\right),\\
&T_{r,1;1,2}(V_0, |a_1(\omega_0)|) = a_1^3(\omega_0) \cdot \frac{\partial V_{1,1}}{\partial x_{4,3}}\left(\mathbf{x}_4^{(0)}\right),\\
&S_{r,1;1,3}(V_0, |a_1(\omega_0)|) = a_1(\omega_0)a_1^{*3}(\omega_0) \cdot \frac{\partial V_{1,1}}{\partial x_{4,9}}\left(\mathbf{x}_4^{(0)}\right) + a_1^{*2}(\omega_0) \cdot V_{1,4}\left(\mathbf{x}_4^{(0)}\right),\\
&T_{r,1;1,3}(V_0, |a_1(\omega_0)|) = a_1^4(\omega_0) \cdot \frac{\partial V_{1,1}}{\partial x_{4,8}}\left(\mathbf{x}_4^{(0)}\right),\\
&S_{r,1;1,4}(V_0, |a_1(\omega_0)|) = a_1(\omega_0)a_1^{*4}(\omega_0) \cdot \frac{\partial V_{1,1}}{\partial x_{4,24}}\left(\mathbf{x}_4^{(0)}\right) + a_1^{*3}(\omega_0) \cdot V_{1,8}\left(\mathbf{x}_4^{(0)}\right),\\
&T_{r,1;1,4}(V_0, |a_1(\omega_0)|) = a_1^5(\omega_0) \cdot \frac{\partial V_{1,1}}{\partial x_{4,23}}\left(\mathbf{x}_4^{(0)}\right).
\end{aligned}$$

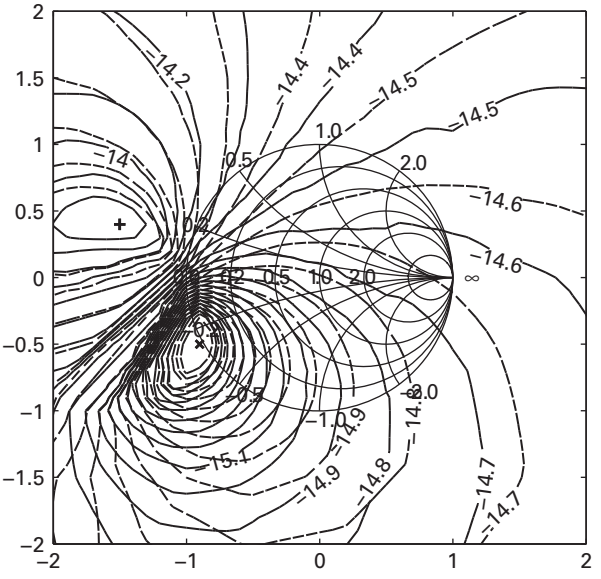


Figure 6.8 Power contour results for the two-tone PHD model.

Having rederived the PHD model from the Volterra expansion, we shall now test its performance for the negative resistance modeled in Section 6.2.5. We first extract a two-tone PHD model since the RTALP experiment considered is only exciting the first and second harmonics. The results shown in Figure 6.8 exhibit a substantial difference between the contour plots obtained from the measured and extracted data compared with Figure 6.6, even though the difference between contour lines is only 0.1 dB.

Next we extract a four-tone PHD model using the measured first, second, third, and fourth harmonics. Much closer results are now obtained in Figure 6.9 for the contour plots obtained for the extracted and measured data. The results obtained with a first-order Volterra expansion for the second, third, and fourth harmonics are now comparable to those obtained in Figure 6.6 using two harmonics but with a Volterra expansion of order eight for the second harmonic. It is as if using more tones has enabled us to reduce the order of the nonlinearity in the model. The source of this effect is related to the impedances provided by the RTALP testbed.

The reflection coefficients exhibited by the RTALP testbed at the third and fourth harmonic have amplitudes on average of 0.019 and 0.075, respectively; thus the testbed is not perfectly matched. As a result the reflected waves $b(3\omega_0)$ and $b(4\omega_0)$ are partially reflected back to the device as incident waves $a(3\omega_0)$ and $a(4\omega_0)$. Further, at high power levels of the second-harmonic signal injected into the device, the fourth-harmonic reflection coefficient is found to briefly increase up to almost unity in a narrow portion of the phase locus. This results from a quasi-resonant nonlinear interaction between the device and the generators. The non-ideal match observed at the third and fourth harmonics increases the apparent nonlinearity of the device-plus-generator system when it

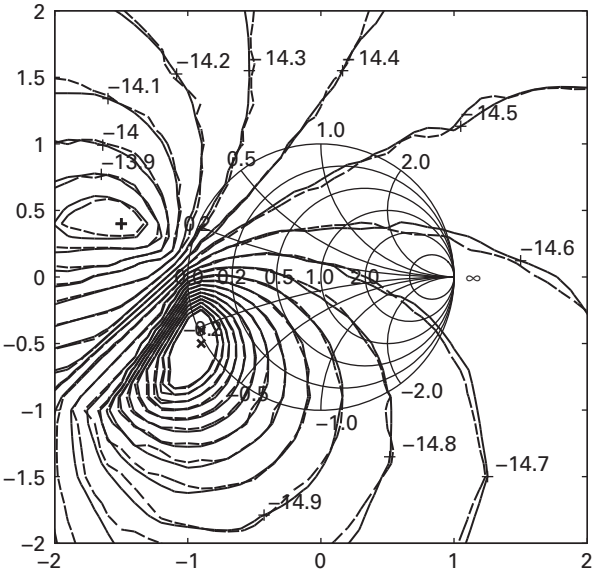


Figure 6.9 Power contour results for the four-tone PHD model.

is modeled using only two harmonics. Indeed, in the two-harmonic model any non-zero incident waves $a(3\omega_0)$ and $a(4\omega_0)$ are absorbed in the device nonlinearity just like any other hidden dependent variables. On the other hand, in the four-harmonic model, these independent variables are simply treated as excitations applied to the device.

This example illustrates (1) the importance of accounting for the non-ideal matching characteristics of the load-pull testbed at all harmonics and (2) the interplay between the nonlinearity order of the behavioral model and the number of tones included in the device model. An effective decrease in the model nonlinearity order can be achieved by accounting for more harmonic tones. This is the strategy of choice since it reduces the model complexity. In mildly nonlinear circuits a linear treatment for the harmonics (PHD approximation) is then sufficient, provided that enough of them are accounted for. Further, the model can then handle different impedance terminations for the harmonic considered.

6.3 Single-band multi-harmonic envelope PA model

Having discussed the case of a SISO system driven by multiple harmonics, let us now consider the case in which the higher harmonics of the input signal are all filtered out and a single tone of radial frequency ω_{RF} is incident on the device. In such a case the multi-harmonic response of the system reduces to

$$b_{\mathrm{out}}(\omega_{\mathrm{RF}}) = f_1(|a_{\mathrm{in}}(\omega_{\mathrm{RF}})|^2) \cdot a_1(\omega_{\mathrm{RF}}),$$
$$b_{\mathrm{out}}(2\omega_{\mathrm{RF}}) = f_2(|a_{\mathrm{in}}(\omega_{\mathrm{RF}})|^2) \cdot a_1^2(\omega_{\mathrm{RF}}),$$

$$
\begin{aligned}
b_{\text{out}}(3\omega_{\text{RF}}) &= f_3(|a_{\text{in}}(\omega_{\text{RF}})|^2)\cdot a_1^3(\omega_{\text{RF}}),\\
b_{\text{out}}(4\omega_{\text{RF}}) &= f_4(|a_{\text{in}}(\omega_{\text{RF}})|^2)\cdot a_1^4(\omega_{\text{RF}}),\\
b_{\text{out}}(5\omega_{\text{RF}}) &= f_5(|a_{\text{in}}(\omega_{\text{RF}})|^2)\cdot a_1^5(\omega_{\text{RF}}),\\
&\cdots\\
b_{\text{out}}(p\omega_{\text{RF}}) &= f_p(|a_{\text{in}}(\omega_{\text{RF}})|^2)\cdot a_1^p(\omega_{\text{RF}}).
\end{aligned}
$$

Up to now we have only considered a CW incident RF signal with no modulation. We shall now discuss the extension to modulated incident signals.

6.3.1 Input signal

Let us assume now that the input RF tone $a_{\text{in}}(t)$ is modulated both in amplitude and in phase:

$$
\begin{aligned}
a_{\text{in}}(t) &= E(t)\cos[\omega_{\text{RF}}t+\theta(t)]\\
&= E(t)\cos[\theta(t)]\cos(\omega_{\text{RF}}t) - E(t)\sin[\theta(t)]\sin(\omega_{\text{RF}}t)\\
&= I(t)\cos(\omega_{\text{RF}}t) - Q(t)\sin(\omega_{\text{RF}}t),
\end{aligned}
$$

with

$$
\begin{aligned}
I(t) &= E(t)\cos[\theta(t)], \qquad & Q(t) &= E(t)\sin[\theta(t)],\\
E^2(t) &= I^2(t)+Q^2(t), \qquad & \theta(t) &= \angle\left[I(t)+jQ(t)\right],
\end{aligned}
$$

where $E(t)$, $\theta(t)$, and ω_{RF} are the time-varying envelope, time-varying phase, and center frequency, respectively, of the input signal $a_{\text{in}}(t)$.

We still assume that no harmonic of ω_{RF} is injected at the input. This is a good assumption if the source driving the input is filtered to remove the higher harmonics and the source is matched so that the harmonics generated by the amplifier at the input are absorbed by the source. The amplifier device considered being nonlinear, multiple harmonics will, however, be generated at the output. For a memoryless nonlinear amplifier we can express the output signal b_{out} in terms of a power series. In the following derivation we will consider an amplifier with nonlinearities up to seventh order:

$$
\begin{aligned}
b_{\text{out}}^{\text{ML}}(t) = a_0 &+ a_1a_{\text{in}}(t) + a_2a_{\text{in}}^2(t) + a_3a_{\text{in}}^3(t) + a_4a_{\text{in}}^4(t)\\
&+ a_5a_{\text{in}}^5(t) + a_6a_{\text{in}}^6(t) + a_7a_{\text{in}}^7(t).
\end{aligned}
\tag{6.36}
$$

6.3.2 Orthogonal Chaillot expansion

Before proceeding with the evaluation of $b_{\text{out}}^{\text{ML}}(t)$ let us define the functions $I_i(t)$ and $Q_i(t)$ with i a positive integer to be used for the orthogonal Chaillot expansion in the frequency domain of $I^i(t)$ and $Q^i(t)$:

$$
I_i(t) = E^i(t)\cos[i\theta(t)], \qquad Q_i(t) = E^i(t)\sin[i\theta(t)]. \tag{6.37}
$$

The even and odd powers of $I(t)$ and $Q(t)$ are then expressed in terms of $I_i(t)$ and $Q_i(t)$ as follows:

$$
\begin{aligned}
I(t) &= I_1(t),\\
I^2(t) &= \frac{1}{2}E^2(t) + \frac{1}{2}E^2(t)\cos[2\theta(t)]\\
&= \frac{1}{2}E^2(t) + \frac{1}{2}I_2(t),\\
I^3(t) &= \frac{3}{4}E^3(t)\cos[\theta(t)] + \frac{1}{4}E^3(t)\cos[3\theta(t)]\\
&= \frac{3}{4}E^2(t)I_1(t) + \frac{1}{4}I_3(t),\\
I^4(t) &= \frac{3}{8}E^4(t) + \frac{1}{2}E^4(t)\cos[2\theta(t)] + \frac{1}{8}E^4(t)\cos[4\theta(t)]\\
&= \frac{3}{8}E^4(t) + \frac{1}{2}E^2(t)I_2(t) + \frac{1}{8}I_4(t),\\
I^5(t) &= \frac{5}{8}E^5(t)\cos[\theta(t)] + \frac{5}{16}E^5(t)\cos[3\theta(t)]\\
&\quad + \frac{1}{16}E^5(t)\cos[5\theta(t)]\\
&= \frac{5}{8}E^4(t)I_1(t) + \frac{5}{16}E^2(t)I_3(t) + \frac{1}{16}I_5(t),\\
I^6(t) &= \frac{10}{32}E^6(t) + \frac{15}{32}E^6(t)\cos[2\theta(t)] + \frac{6}{32}E^6(t)\cos[4\theta(t)]\\
&\quad + \frac{1}{32}E^6(t)\cos[6\theta(t)]\\
&= \frac{10}{32}E^6(t) + \frac{15}{32}E^4(t)I_2(t) + \frac{6}{32}E^2(t)I_4(t) + \frac{1}{32}I_6(t),\\
I^7(t) &= \frac{35}{64}E^7(t)\cos[\theta(t)] + \frac{21}{64}E^7(t)\cos[3\theta(t)]\\
&\quad + \frac{7}{64}E^7(t)\cos[5\theta(t)] + \frac{1}{64}E^7(t)\cos[7\theta(t)]\\
&= \frac{35}{64}E^6(t)I_1(t) + \frac{21}{64}E^4(t)I_3(t) + \frac{7}{64}E^2(t)I_5(t) + \frac{1}{64}I_7(t),
\end{aligned}
$$

and

$$
\begin{aligned}
Q(t) &= Q_1(t),\\
Q^3(t) &= \frac{3}{4}E^2(t)Q_1(t) - \frac{1}{4}Q_3(t),\\
Q^5(t) &= \frac{5}{8}E^4(t)Q_1(t) - \frac{5}{16}E^2(t)Q_3(t) + \frac{1}{16}Q_5(t),\\
Q^7(t) &= \frac{35}{64}E^6(t)Q_1(t) - \frac{21}{64}E^4(t)Q_3(t) + \frac{7}{64}E^2(t)Q_5(t) - \frac{1}{64}Q_7(t).
\end{aligned}
$$

The Chaillot orthogonal components $I_i(t)$ and $Q_i(t)$ can then be directly calculated in terms of $I(t)$ and $Q(t)$ using

$$
\begin{aligned}
X_0(t) &= 1,\\
I_1(t) &= I(t),\\
I_2(t) &= 2I^2(t) - E^2(t),\\
I_3(t) &= 4I^3(t) - 3E^2(t)I(t),\\
I_4(t) &= 8I^4(t) - 4E^2(t)I_2(t) - 3E^4(t)\\
&= 8I^4(t) - 8E^2(t)I^2(t) + E^4(t),\\
I_5(t) &= 16I^5(t) - 5E^2(t)I_3(t) - 10E^4(t)I_1(t)\\
&= 16I^5(t) - 20E^2(t)I^3(t) + 5E^4(t)I(t),\\
I_6(t) &= 32I^6(t) - 6E^2(t)I_4(t) - 15E^4(t)I_2(t) - 10E^6(t)\\
&= 32I^6(t) - 48E^2(t)I^4(t) + 18E^4(t)I^2(t) - E^6(t),\\
I_7(t) &= 64I^7(t) - 7E^2(t)I_5(t) - 21E^4(t)I_3(t) - 35E^6(t)I_1(t)\\
&= 64I^7(t) - 112E^2(t)I^5(t) + 56E^4(t)I^3(t) - 7E^6(t)I(t),
\end{aligned}
$$

and

$$
\begin{aligned}
Q_0(t) &= 0,\\
Q_1(t) &= Q(t),\\
Q_2(t) &= E^2(t)\sin[2\theta(t)] = 2I_1(t)Q_1(t),\\
Q_3(t) &= -4Q^3(t) + 3E^2(t)Q(t),\\
Q_4(t) &= E^4(t)\sin[4\theta(t)] = 2I_2(t)Q_2(t),\\
Q_5(t) &= 16Q^5(t) + 5E^2(t)Q_3(t) - 10E^4(t)Q_1(t)\\
&= 16Q^5(t) - 20E^2(t)Q^3(t) + 5E^4(t)Q(t),\\
Q_6(t) &= E^6(t)\sin[6\theta(t)] = 2I_3(t)Q_3(t),\\
Q_7(t) &= -64Q^7(t) + 7E^2(t)Q_5(t) - 21E^4(t)Q_3(t) + 35E^6(t)Q_1(t)\\
&= -64Q^7(t) + 112E^2(t)Q^5(t) - 56E^4(t)Q^3(t) + 7E^6(t)Q(t).
\end{aligned}
$$

Note that, in the limit $E^2(t) = 1$, the Chaillot functions $I_i(t)$ reduce to the Chebyshev polynomials of the first kind, and $Q_i(t)$ reduce to $(-1)^{\text{floor}[i/2]}$ of the Chebyshev polynomials of the first kind, for i odd.

Similarly, for time-varying envelope $E^2(t)$ the Chaillot functions $I_i(t)$ of arbitrary order can be obtained using the following recurrence equation:

$$
I_i(t) = \left\{ \begin{array}{ll} 1, & \text{for } i = 0 \\ I(t), & \text{for } i = 1 \\ 2I(t)I_{i-1}(t) - E^2(t)I_{i-2}(t), & i > 1 \end{array} \right\} = \mathcal{T}_i\left[I(t)\right], \tag{6.38}
$$

where $\mathcal{T}_i[I]$ are generalized Chebyshev functions of the first kind. The Chaillot functions $Q_i(t)$ are calculated as

$$
Q_i(t) = \left\{ \begin{array}{ll} 0, & \text{for } i = 0 \\ Q(t), & \text{for } i = 1 \\ (-1)^{\text{floor}[i/2]}\mathcal{T}_i\left[Q(t)\right], & \text{for } i \text{ odd} \\ 2I_{i/2}(t)Q_{i/2}(t), & \text{for } i \text{ even.} \end{array} \right. \tag{6.39}
$$

6.3.3 Memoryless nonlinear system modeling

To analyze up to the seventh order a memoryless nonlinear system described by Equation (6.36), we need to evaluate the following powers of a_{in}:

$$
\begin{aligned}
a_{\text{in}}(t) &= E(t)\cos[\omega_{\text{RF}}t + \theta(t)] \\
&= I_1(t)\cos(\omega_{\text{RF}}t) - Q_1(t)\sin(\omega_{\text{RF}}t), \\
a_{\text{in}}^2(t) &= \{E(t)\cos[\omega_{\text{RF}}t + \theta(t)]\}^2 \\
&= \frac{1}{2}E^2(t) + \frac{1}{2}\{I_2(t)\cos(2\omega_{\text{RF}}t) - Q_2(t)\sin(2\omega_{\text{RF}}t)\}, \\
a_{\text{in}}^3(t) &= \{E(t)\cos[\omega_{\text{RF}}t + \theta(t)]\}^3 \\
&= \frac{3}{4}E^2(t)\{I_1(t)\cos(\omega_{\text{RF}}t) - Q_1(t)\sin(\omega_{\text{RF}}t)\} \\
&\quad + \frac{1}{4}\{I_3(t)\cos(3\omega_{\text{RF}}t) - Q_3(t)\sin(3\omega_{\text{RF}}t)\}, \\
a_{\text{in}}^4(t) &= \{E(t)\cos[\omega_{\text{RF}}(t) + \theta(t)]\}^4 \\
&= \frac{3}{8}E^4(t) \\
&\quad + \frac{1}{2}E^2(t)\{I_2(t)\cos(2\omega_{\text{RF}}t) - Q_2(t)\sin(2\omega_{\text{RF}}t)\} \\
&\quad + \frac{1}{8}\{I_4(t)\cos(4\omega_{\text{RF}}t) - Q_4(t)\sin(4\omega_{\text{RF}}t)\}, \\
a_{\text{in}}^5(t) &= \{E(t)\cos[\omega_{\text{RF}}(t) + \theta(t)]\}^5 \\
&= \frac{5}{8}E^4(t)\{I_1(t)\cos(\omega_{\text{RF}}t) - Q_1(t)\sin(\omega_{\text{RF}}t)\} \\
&\quad + \frac{5}{16}E^2(t)\{I_3(t)\cos(3\omega_{\text{RF}}t) - Q_3(t)\sin(3\omega_{\text{RF}}t)\} \\
&\quad + \frac{1}{16}\{I_5(t)\cos(5\omega_{\text{RF}}t) - Q_5(t)\sin(5\omega_{\text{RF}}t)\}, \\
a_{\text{in}}^6(t) &= \{E(t)\cos[\omega_{\text{RF}}(t) + \theta(t)]\}^6 \\
&= \frac{10}{32}E^6(t) \\
&\quad + \frac{15}{32}E^4(t)\{I_2(t)\cos(2\omega_{\text{RF}}t) - Q_2(t)\sin(2\omega_{\text{RF}}t)\} \\
&\quad + \frac{6}{32}E^2(t)\{I_4(t)\cos(4\omega_{\text{RF}}t) - Q_4(t)\sin(4\omega_{\text{RF}}t)\} \\
&\quad + \frac{1}{32}\{I_6(t)\cos(6\omega_{\text{RF}}t) - Q_6(t)\sin(6\omega_{\text{RF}}t)\}, \\
a_{\text{in}}^7(t) &= \{E(t)\cos[\omega_{\text{RF}}(t) + \theta(t)]\}^7 \\
&= \frac{35}{64}E^6(t)\{I_1(t)\cos(\omega_{\text{RF}}t) - Q_1(t)\sin(\omega_{\text{RF}}t)\} \\
&\quad + \frac{21}{64}E^4(t)\{I_3(t)\cos(3\omega_{\text{RF}}t) - Q_3(t)\sin(3\omega_{\text{RF}}t)\}
\end{aligned}
$$

$$+\frac{7}{64}E^2(t)\{I_5(t)\cos(5\omega_{\mathrm{RF}}t)-Q_5(t)\sin(5\omega_{\mathrm{RF}}t)\}$$
$$+\frac{1}{64}\{I_7(t)\cos(7\omega_{\mathrm{RF}}t)-Q_7(t)\sin(7\omega_{\mathrm{RF}}t)\}.$$

To infer the form of the modeling equations needed for such a nonlinear system, it is very useful to sort the nonlinear terms generated in terms of frequencies rather than in terms of the order of the nonlinearity generating them. The final output of the seventh-order memoryless nonlinear system is

$$\begin{aligned}
b_{\mathrm{out}}^{\mathrm{ML}}(t) &= a_0 + a_1a_{\mathrm{in}}(t) + a_2a_{\mathrm{in}}^2(t) + a_3a_{\mathrm{in}}^3(t) + a_4a_{\mathrm{in}}^4(t) + a_5a_{\mathrm{in}}^5(t) + a_6a_{\mathrm{in}}^6(t) \\
&\quad + a_7a_{\mathrm{in}}^7(t) \\
&= f_0(E^2) \\
&\quad + f_1(E^2)[I_1(t)\cos(\omega_{\mathrm{RF}}t) - Q_1(t)\sin(\omega_{\mathrm{RF}}t)] \\
&\quad + f_2(E^2)[I_2(t)\cos(2\omega_{\mathrm{RF}}t) - Q_2(t)\sin(2\omega_{\mathrm{RF}}t)] \\
&\quad + f_3(E^2)[I_3(t)\cos(3\omega_{\mathrm{RF}}t) - Q_3(t)\sin(3\omega_{\mathrm{RF}}t)] \\
&\quad + f_4(E^2)[I_4(t)\cos(4\omega_{\mathrm{RF}}t) - Q_4(t)\sin(4\omega_{\mathrm{RF}}t)] \\
&\quad + f_5(E^2)[I_5(t)\cos(5\omega_{\mathrm{RF}}t) - Q_5(t)\sin(5\omega_{\mathrm{RF}}t)] \\
&\quad + f_6(E^2)[I_6(t)\cos(6\omega_{\mathrm{RF}}t) - Q_6(t)\sin(6\omega_{\mathrm{RF}}t)] \\
&\quad + f_7(E^2)[I_7(t)\cos(7\omega_{\mathrm{RF}}t) - Q_7(t)\sin(7\omega_{\mathrm{RF}}t)],
\end{aligned}$$

with the function $f_i(E^2)$ given by

$$\begin{aligned}
f_0(E^2) &= a_0 + \frac{1}{2}a_2E^2 + \frac{3}{8}a_4E^4 + \frac{10}{32}a_6E^6, \\
f_1(E^2) &= a_1 + \frac{3}{4}a_3E^2 + \frac{5}{8}a_5E^4 + \frac{35}{64}a_7E^6, \\
f_2(E^2) &= \frac{1}{2}a_2 + \frac{1}{2}a_4E^2 + \frac{15}{32}a_6E^4, \\
f_3(E^2) &= \frac{1}{4}a_3 + \frac{5}{16}a_5E^2 + \frac{21}{64}a_7E^4, \\
f_4(E^2) &= \frac{1}{8}a_4 + \frac{6}{32}a_6E^2, \\
f_5(E^2) &= \frac{1}{16}a_5 + \frac{7}{64}a_7E^2, \\
f_6(E^2) &= \frac{1}{32}a_6, \\
f_7(E^2) &= \frac{1}{64}a_7.
\end{aligned}$$

From the preceding derivation we infer that the output of a memoryless nonlinear system represented by an nth-order power series can be modeled in general as follows:

$$b_{\mathrm{out}}^{\mathrm{ML}}(t) = \sum_{i=0}^{n} f_i(E^2)[I_i(t)\cos(i\omega_{\mathrm{RF}}t) - Q_i(t)\sin(i\omega_{\mathrm{RF}}t)], \tag{6.40}$$

$$f_i(E^2) = \sum_{k=0}^{[n-i+\text{even}(i)]/2} \alpha_{i,2k+i} E^{2k}, \tag{6.41}$$

where n is the highest order of the nonlinear terms, $i = 0, 1, 2, \ldots, n$, k is an integer, and even(i) is 1 for i even and 0 otherwise; $\alpha_{i,2k+i}$ gives the contribution weight of the $(2k + i)$th-order nonlinearities to the ith-harmonic band ($i \times \omega_{\text{RF}}$).

6.3.4 Quasi-memoryless nonlinear system modeling

The models expressed by Equation (6.40) are applicable only to memoryless systems because the functions $f_i(E^2)$ are dependent only on the instantaneous value of the envelope square E^2 of the baseband input pair $[I(t), Q(t)]$ and no phase shift is used to account for the system group delay. To account for quasi-memoryless effects [10] it is necessary to introduce an independent phase shift ϕ_i for each harmonic $i \times \omega_{\text{RF}}$. In the general case these phase shifts ϕ_i are also functions of the instantaneous value of the envelope square E^2 (the AM/PM effect). By adding this phase contribution to Equation (6.40) of the memoryless nonlinear system considered so far, we obtain the following modeling equations for a quasi-memoryless (QML) nonlinear amplifier:

$$b_{\text{out}}^{\text{QML}}(t) = \sum_{i=0}^{n} f_i(E^2) \left\{ I_i(t)\cos\left[i\omega_{\text{RF}}t + \phi_i(E^2)\right] - Q_i(t)\sin\left[i\omega_{\text{RF}}t + \phi_i(E^2)\right] \right\}. \tag{6.42}$$

On expanding $\cos\left[i\omega_{\text{RF}}t + \phi_i(E^2)\right]$ and $\sin\left[i\omega_{\text{RF}}t + \phi_i(E^2)\right]$ and factorizing the terms $\cos(i\omega_{\text{RF}}t)$ and $\sin(i\omega_{\text{RF}}t)$, respectively, the modeling equations for the QML nonlinear system can be rewritten as

$$b_{\text{out}}^{\text{QML}}(t) = \sum_{i=0}^{n} \left\{ \cos(i\omega_{\text{RF}}t)\left[I_i(t)g_i(E^2) - Q_i(t)h_i(E^2)\right] - \sin(i\omega_{\text{RF}}t)\left[I_i(t)h_i(E^2) + Q_i(t)g_i(E^2)\right] \right\}, \tag{6.43}$$

where we introduced the functions $g_i(E^2)$ and $h_i(E^2)$:

$$g_i(E^2) = f_i(E^2)\cos\left[\phi_i(E^2)\right], \tag{6.44}$$

$$h_i(E^2) = f_i(E^2)\sin\left[\phi_i(E^2)\right]. \tag{6.45}$$

Now, by defining the nonlinearly scaled and phase-shifted I_i' and Q_i' modulation terms

$$\begin{bmatrix} I_i'(t) \\ Q_i'(t) \end{bmatrix} = f_i(E^2) \begin{bmatrix} \cos\phi_i(E^2) & -\sin\phi_i(E^2) \\ \sin\phi_i(E^2) & \cos\phi_i(E^2) \end{bmatrix} \times \begin{bmatrix} I_i(t) \\ Q_i(t) \end{bmatrix} = \begin{bmatrix} g_i(E^2) & -h_i(E^2) \\ h_i(E^2) & g_i(E^2) \end{bmatrix} \times \begin{bmatrix} I_i(t) \\ Q_i(t) \end{bmatrix}, \tag{6.46}$$

we obtain the following compact equation:

$$b_{\text{out}}^{\text{QML}}(t) = \sum_{i=0}^{n} \left[I_i'(t)\cos(i\omega_{\text{RF}}t) - Q_i'(t)\sin(i\omega_{\text{RF}}t) \right]. \tag{6.47}$$

6.3.5 Power-series expansion

From Equations (6.44) and (6.45), it is seen that the nonlinearities of the system originate from two nonlinear sources: (1) the amplitude-distortion nonlinearity represented by the functions $f_i(E^2)$ and (2) the phase-distortion nonlinearity represented by the functions $\phi_i(E^2)$. For hard nonlinearity these functions are best represented by splines such as the B-spline representation mentioned in Chapter 3.

On the other hand, for low enough input signal power, the functions $g_i(E^2)$ and $h_i(E^2)$ can be expanded in a Taylor series like in Equation (6.41):

$$g_i(E^2) = \sum_{k=0}^{[n-i+\text{even}(i)]/2} \alpha_{i,2k+i} \cdot \cos\varphi_{i,2k+i} \cdot E^{2k}, \tag{6.48}$$

$$h_i(E^2) = \sum_{k=0}^{[n-i+\text{even}(i)]/2} \alpha_{i,2k+i} \cdot \sin\varphi_{i,2k+i} \cdot E^{2k}. \tag{6.49}$$

So, according to Equations (6.48) and (6.49), for a seventh-odd-order QML system in the weakly nonlinear regime, we have

$$\begin{aligned}
g_0(E^2) &= \alpha_{0,0}\cos\varphi_{0,0} + \alpha_{0,2}\cos\varphi_{0,2}\,E^2 + \alpha_{0,4}\cos\varphi_{0,4}\,E^4 + \alpha_{0,6}\cos\varphi_{0,6}\,E^6,\\
g_1(E^2) &= \alpha_{1,1}\cos\varphi_{1,1} + \alpha_{1,3}\cos\varphi_{1,3}\,E^2 + \alpha_{1,5}\cos\varphi_{1,5}\,E^4 + \alpha_{1,7}\cos\varphi_{1,7}\,E^6,\\
g_2(E^2) &= \alpha_{2,2}\cos\varphi_{2,2} + \alpha_{2,4}\cos\varphi_{2,4}\,E^2 + \alpha_{2,6}\cos\varphi_{2,6}\,E^4,\\
g_3(E^2) &= \alpha_{3,3}\cos\varphi_{3,3} + \alpha_{3,5}\cos\varphi_{3,5}\,E^2 + \alpha_{3,7}\cos\varphi_{3,7}\,E^4,\\
g_4(E^2) &= \alpha_{4,4}\cos\varphi_{4,4} + \alpha_{4,6}\cos\varphi_{4,6}\,E^2,\\
g_5(E^2) &= \alpha_{5,5}\cos\varphi_{5,5} + \alpha_{5,7}\cos\varphi_{5,7}\,E^2,\\
g_6(E^2) &= \alpha_{6,6}\cos\varphi_{6,6},\\
g_7(E^2) &= \alpha_{7,7}\cos\varphi_{7,7},
\end{aligned}$$

and

$$\begin{aligned}
h_0(E^2) &= \alpha_{0,0}\sin\varphi_{0,0} + \alpha_{0,2}\sin\varphi_{0,2}\,E^2 + \alpha_{0,4}\sin\varphi_{0,4}\,E^6 + \alpha_{0,6}\sin\varphi_{0,6}\,E^6,\\
h_1(E^2) &= \alpha_{1,1}\sin\varphi_{1,1} + \alpha_{1,3}\sin\varphi_{1,3}\,E^2 + \alpha_{1,5}\sin\varphi_{1,5}\,E^4 + \alpha_{1,7}\sin\varphi_{1,7}\,E^6,\\
h_2(E^2) &= \alpha_{2,2}\sin\varphi_{2,2} + \alpha_{2,4}\sin\varphi_{2,4}\,E^2 + \alpha_{2,6}\sin\varphi_{2,6}\,E^4,\\
h_3(E^2) &= \alpha_{3,3}\sin\varphi_{3,3} + \alpha_{3,5}\sin\varphi_{3,5}\,E^2 + \alpha_{3,7}\sin\varphi_{3,7}\,E^4,\\
h_4(E^2) &= \alpha_{4,4}\sin\varphi_{4,4} + \alpha_{4,6}\sin\varphi_{4,6}\,E^2,\\
h_5(E^2) &= \alpha_{5,5}\sin\varphi_{5,5} + \alpha_{5,7}\sin\varphi_{5,7}\,E^2,\\
h_6(E^2) &= \alpha_{6,6}\sin\varphi_{6,6},\\
h_7(E^2) &= \alpha_{7,7}\sin\varphi_{7,7}.
\end{aligned}$$

6.3.6 Multi-path model partitioning

A more complex nonlinear model can sometimes advantageously be obtained by superposing several models in parallel, which can then be represented by different power-series expansions. Let us motivate the physical origin for such multi-path modeling.

As we can infer from our derivations in Section 6.3.3, a memoryless nonlinear term X^n contributes 2^{n-1} nonlinear sub-paths. For example, for $n = 5$ the coefficients 10/16, 5/16 and 1/16 weighting $\cos(\theta)$, $\cos(3\theta)$ and $\cos(5\theta)$, respectively, which sum to $16/16 = 1$, are associated with 16 paths. But, more importantly, several paths can contribute to the same nonlinearity at the harmonic $i \times \theta$. In the QML case, each path contributing to the same nonlinearity is associated with not only a different amplitude but also a different phase, so that the final coefficient for a specific nonlinearity is the summation of the phasor of each path.

In some modeling situations these paths can be approximately independent and the behavioral model can be beneficially presented using a summation over multiple paths P:

$$g_i(E^2) = \sum_{p=1}^{P} g_i^{(p)}(E^2) \qquad \text{and} \qquad h_i(E^2) = \sum_{p=1}^{P} h_i^{(p)}(E^2).$$

In Chapter 8 we shall give an example of a nonlinear modulator switching from the weak nonlinear regime to the hard nonlinear regime and demonstrate the benefit from such multiple-path modeling.

6.3.7 Time-selective single-band multi-harmonic envelope PA model

The multi-harmonic model presented in the previous section is a piecewise QML model; that is, only AM/AM and AM/PM effects have been accounted for each harmonic. To account for dynamic memory effects, a combined frequency-selective and time-selective model can be realized by using the memory polynomial approximation [7] [8]. This approximation is equivalent to setting all the delays equal to the same values τ throughout the Volterra expansion, effectively retaining only the diagonal terms. For example, in the case of the third-order Volterra kernel, in Equation (6.2) the three-dimensional integration collapses into a one-dimensional integration with $\tau_1 = \tau_2 = \tau_3 = \tau$. For the sake of discrete time processing, the integration can be replaced by a summation of the delay τ_p, yielding at the fundamental frequency, for a system of order n,

$$y(m) = \sum_{k=0}^{n} \sum_{p=0}^{P} a_{kp} \cdot x(m-p)|x(m-p)|^k,$$

with $x(k) = I_1(k) + jQ_1(k)$ the input signal, $y(k) = I_1'(k) + jQ_1'(k)$ the output signal, P the memory depth, and a_{kp} complex coefficients. The output is then the sum of different nonlinear models driven by different delayed input signals.

To generalize this memory polynomial model to multi-harmonic output signals we switch back to the matrix *I–Q* representation. The output I_i' and Q_i' modulation terms for each harmonic i are now the summations of different components that are not only nonlinearly scaled and phase-shifted but also time-shifted:

$$\begin{bmatrix} I_i'(t) \\ Q_i'(t) \end{bmatrix} = \sum_p \left\{ \begin{bmatrix} g_{i,p}[E^2(t-\tau_p)] & -h_{i,p}[E^2(t-\tau_p)] \\ h_{i,p}[E^2(t-\tau_p)] & g_{i,p}[E^2(t-\tau_p)] \end{bmatrix} \times \begin{bmatrix} I_i(t-\tau_p) \\ Q_i(t-\tau_p) \end{bmatrix} \right\}. \tag{6.50}$$

Note that the different delays τ_p used need not necessarily be equally distributed. But in practice we usually have $\tau_p = p\tau_S$ with $1/\tau_S = f_S$ the sampling frequency required to address memory of a specific order O. For example, if the baseband bandwidth is B and the intermodulation nonlinearity order is O, the minimum sampling frequency required is $f_S = 2 \times O \times B$.

Note that the RF output accounting for the memory effect (ME) is still given by the equations

$$b_{\text{out}}^{\text{ME}}(t) = \sum_{i=0}^{n} \left[I_i'(t)\cos(i\omega_{\text{RF}}t) - Q_i'(t)\sin(i\omega_{\text{RF}}t) \right]. \tag{6.51}$$

This constitutes a generalization of the memory-polynomial approximation to multiple harmonics. This multi-harmonic theory will find application in Chapter 8 for the linearization of a single-sideband modulator by removing the spurious harmonics of the digital IF.

An example of this envelope modeling is given in Figures 6.10 and 6.11 for the fundamental frequency for an 80-W GaN Doherty PA operating at average output power

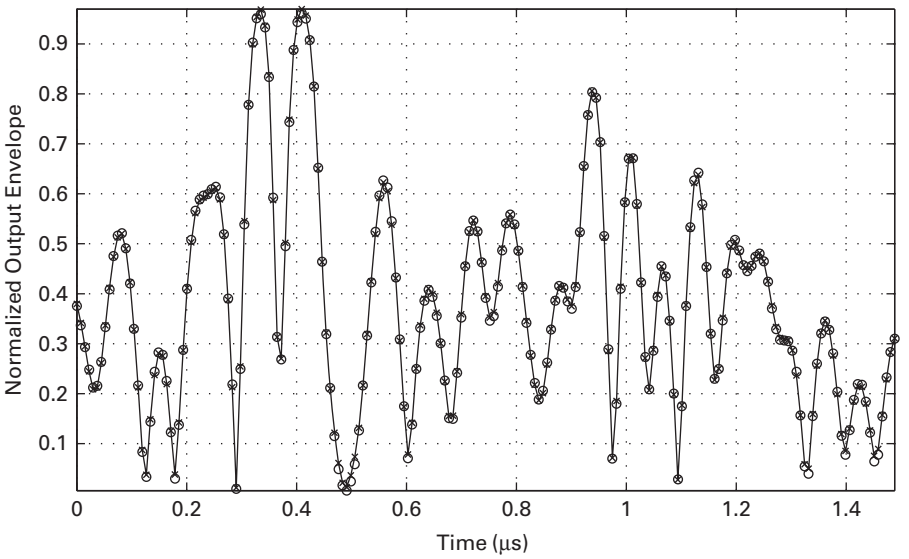


Figure 6.10 The output envelope for an 80-W GaN Doherty amplifier excited by a two-carrier WiMAX signal. Superposed are the measured (circles) and predicted (crosses and line) data.

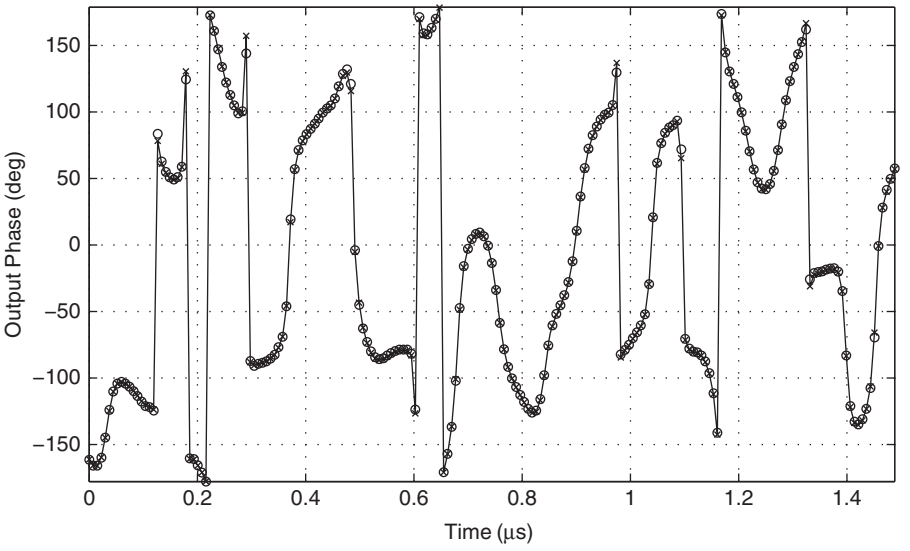


Figure 6.11 The output phase for an 80-W GaN Doherty amplifier excited by a two-carrier WiMAX signal.

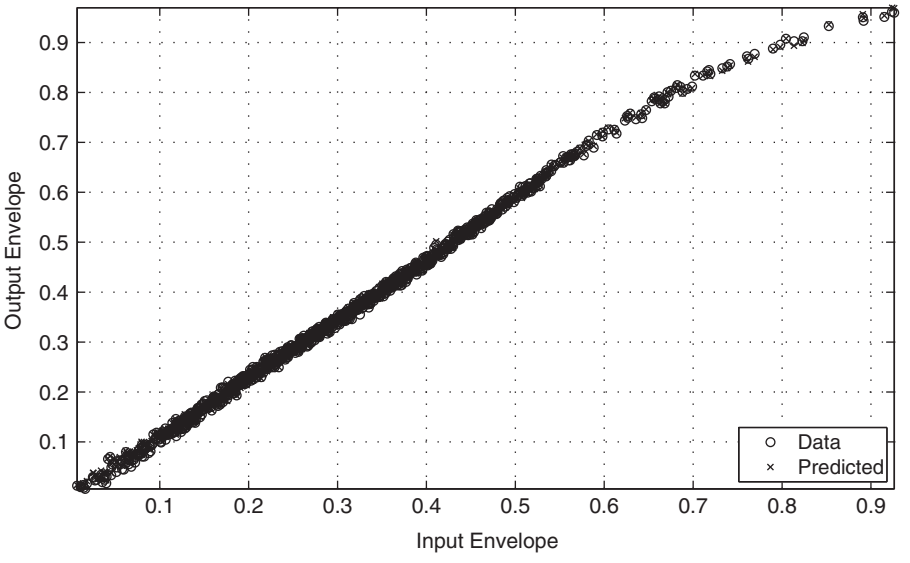


Figure 6.12 The AM/AM conversion characteristic of an 80-W GaN Doherty amplifier for a two-carrier WiMAX signal. Superposed are the measured (circles) and predicted (crosses) data.

48.5 dBm. The model was extracted for one WiMAX symbol and used to predict a different WiMAX symbol. B-spline representation of order 5 with eight intervals together with a memory depth of 5 was used.

Although the agreement between measured and predicted data appears reasonable, a more revealing inspection is provided by the AM/AM and AM/PM results shown in Figures 6.12 and 6.13. The non-zero thickness of the AM/AM and AM/PM curves

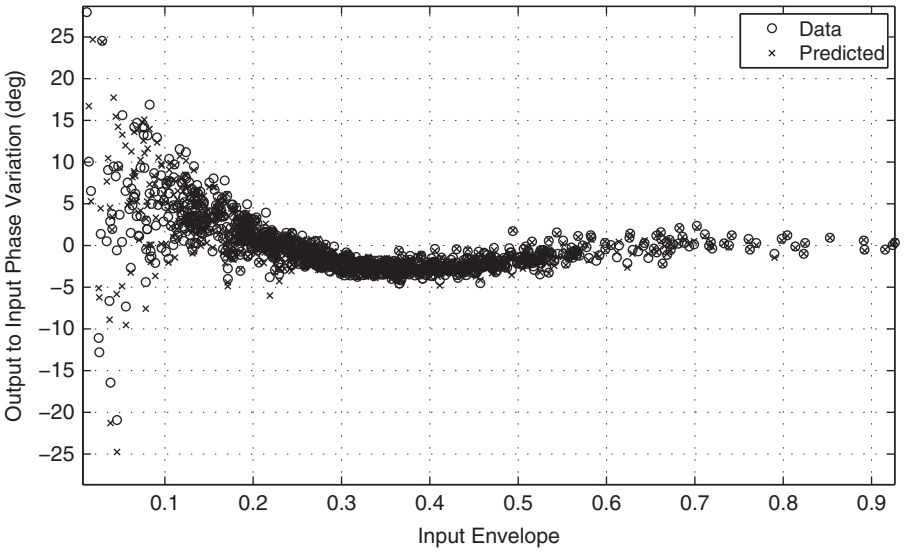


Figure 6.13 The AM/PM conversion characteristic of an 80-W GaN Doherty amplifier for a two-carrier WiMAX signal. Superposed are the measured (circles) and predicted (crosses) data.

constitutes a direct observation of memory effects. It would be highly desirable if the predicted data (crosses) were tracking the measured data (circles). However, the model is seen to yield a large phase error for weak input power levels while yielding more accurate phase at high input power levels. This most probably originates from the weaker output signal being more sensitive to memory effects induced by previous stronger excitations. The RMS error between measured and predicted data is typically found to be 1% when using this B-spline model and the B-spline parameters used here. The spectrum of the data is found to be semi-quantatively predicted by the model as is shown in Figure 6.14. Improved fitting of the data can be obtained in the extraction by accounting for non-diagonal elements. However, the predicted data (reported above) for different symbols were found to degrade for the particular PA considered. The memory-polynomial approximation is therefore a preferable choice for this PA even though the model performance is not fully satisfactory. A more complex model topology featuring additional hidden variables/layers is then necessary for an improved modeling of the memory effects associated with the multistage system. In general, the wider the modulation bandwidth the more challenging the modeling of the memory effects.

6.4 Two-band fundamental envelope PA model

Let us consider now the case when a two-tone excitation $a_1(\omega_1)$ and $a_1(\omega_2)$ of frequencies ω_1 and ω_2, respectively, is applied at port 1 of an amplifier.

Let us initially assume the nonlinearities of the amplifier to be well represented by a third-order Volterra system. For the two-tone excitation the output b_2 at port 2 of the amplifier can be verified to be of the following form (see Figure 6.15):

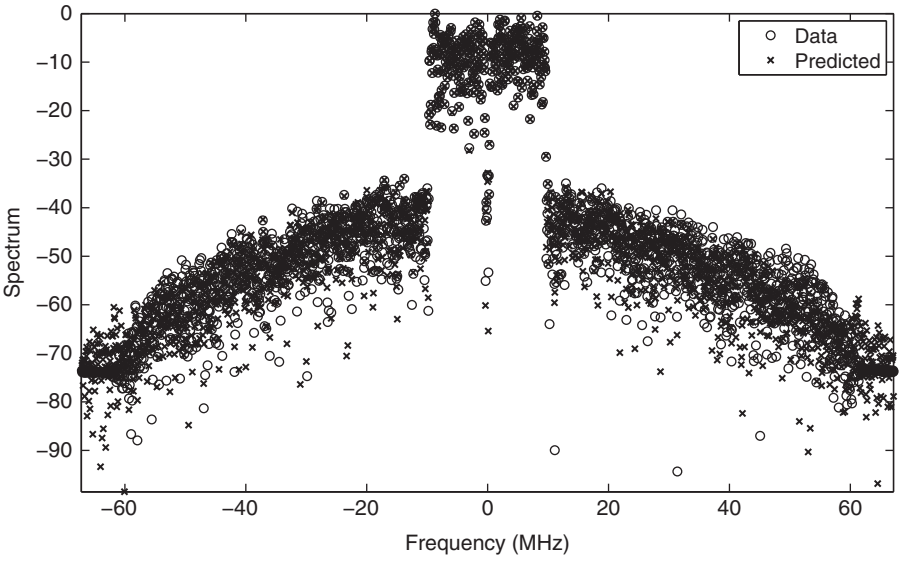


Figure 6.14 The output spectrum of an 80-W GaN Doherty amplifier for a two-carrier WiMAX signal. Superposed are the measured (circles) and predicted (crosses) data.

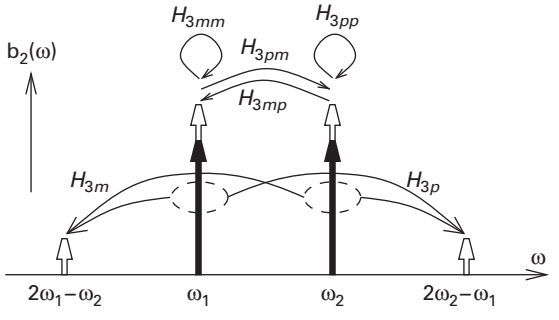


Figure 6.15 Third-order intermodulation for two-tone excitation. (From [11] with permission, ©2005 IEEE.)

$$
\begin{aligned}
b_2(2\omega_1-\omega_2) &= H_{3m}a_1^2(\omega_1)a_1^*(\omega_2),\\
b_2(\omega_1) &= H_{1m}a_1(\omega_1)+H_{3mm}a_1(\omega_1)|a_1(\omega_1)|^2+H_{3mp}a_1(\omega_1)|a_1(\omega_2)|^2,\\
b_2(\omega_2) &= H_{1p}a_1(\omega_2)+H_{3pm}a_1(\omega_2)|a_1(\omega_1)|^2+H_{3pp}a_1(\omega_2)|a_1(\omega_2)|^2,\\
b_2(2\omega_2-\omega_1) &= H_{3p}a_1^2(\omega_2)a_1^*(\omega_1).
\end{aligned}
$$

This output features the two desired tones ω_1 and ω_2 plus two intermodulation tones, $2\omega_1-\omega_2$ and $2\omega_2-\omega_1$. The complex coefficients H_{1m}, H_{1p}, H_{3m}, H_{3p}, H_{3mm}, H_{3mp}, H_{3pm}, and H_{3pp} which are calculated from third-order Volterra series can be measured using a nonlinear network analyzer (NVNA). For a memoryless PA with $y(t)=T_1x(t)+T_3x^3(t)$ we have $H_{1m}=H_{1p}=T_1$ and $H_{3m}=H_{3p}=H_{3mm}=H_{3pp}=3T_3$ and $H_{3mp}=H_{3pm}=6T_3$.

The Volterra series has been extended to larger input power for stochastic input signals with constant average power levels by Wiener [4]. Alternatively, for deterministic signals with large input power, one can also generalize the Volterra formalism by making the H coefficients introduced above power dependent, as discussed in the following sections.

6.4.1 Nonlinear power-amplifier characterization with NVNA

The nonlinear characterization of the PA can be performed using an NVNA as shown in Figure 6.16. With an NVNA we can measure the amplitude and phase of the incident and transmitted, periodically modulated waves at the fundamental and harmonics [12].

In Figure 6.16, an LSNA was used to characterize the third-order intermodulation response of a class-AB LDMOSFET 10-W PA at 895 MHz for a two-tone excitation [11]. The LSNA is used to measure both the amplitude and the phase of the two-tone RF excitation $a_1(\omega_1)$ and $a_1(\omega_2)$ incident on port 1, as well as the amplitude and phase of the transmitted intermodulation RF signals $b_2(2\omega_1 - \omega_2)$, $b_2(\omega_1)$, $b_2(\omega_2)$, and $b_2(2\omega_2 - \omega_1)$ transmitted to port 2. Using these intermodulation signals, we next calculated the *generalized* Volterra coefficients H_{3m} and H_{3p} defined as

$$H_{3m}(\omega_1, \omega_2, |a_1(\omega_1)|^2, |a_1(\omega_2)|^2) = \frac{b_2(2\omega_1 - \omega_2)}{a_1^2(\omega_1)a_1^*(\omega_2)}, \tag{6.52}$$

$$H_{3p}(\omega_1, \omega_2, |a_1(\omega_1)|^2, |a_1(\omega_2)|^2) = \frac{b_2(2\omega_2 - \omega_1)}{a_1^*(\omega_1)a_1^2(\omega_2)}. \tag{6.53}$$

Note that these coefficients are found to be reproducible measurement after measurement [11].

The amplitude and phase of these coefficients are plotted in Figure 6.17 respectively as a function of tone spacing $\omega_m = \omega_2 - \omega_1$ and input power $|\, a_1(\omega_1)\,|^2 = |\, a_1(\omega_2)\,|^2$ from -4 to 6 dBm. As shown in Figure 6.17, the variation of H_{3m} and H_{3p} as a function

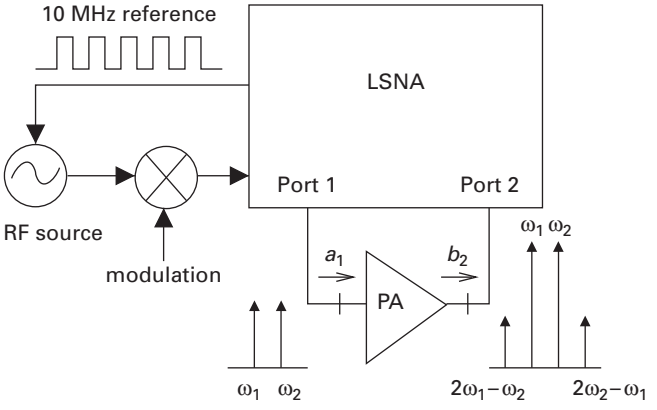


Figure 6.16 A large-signal testbed used for nonlinear characterization. (From [11] with permission, ©2005 IEEE.)

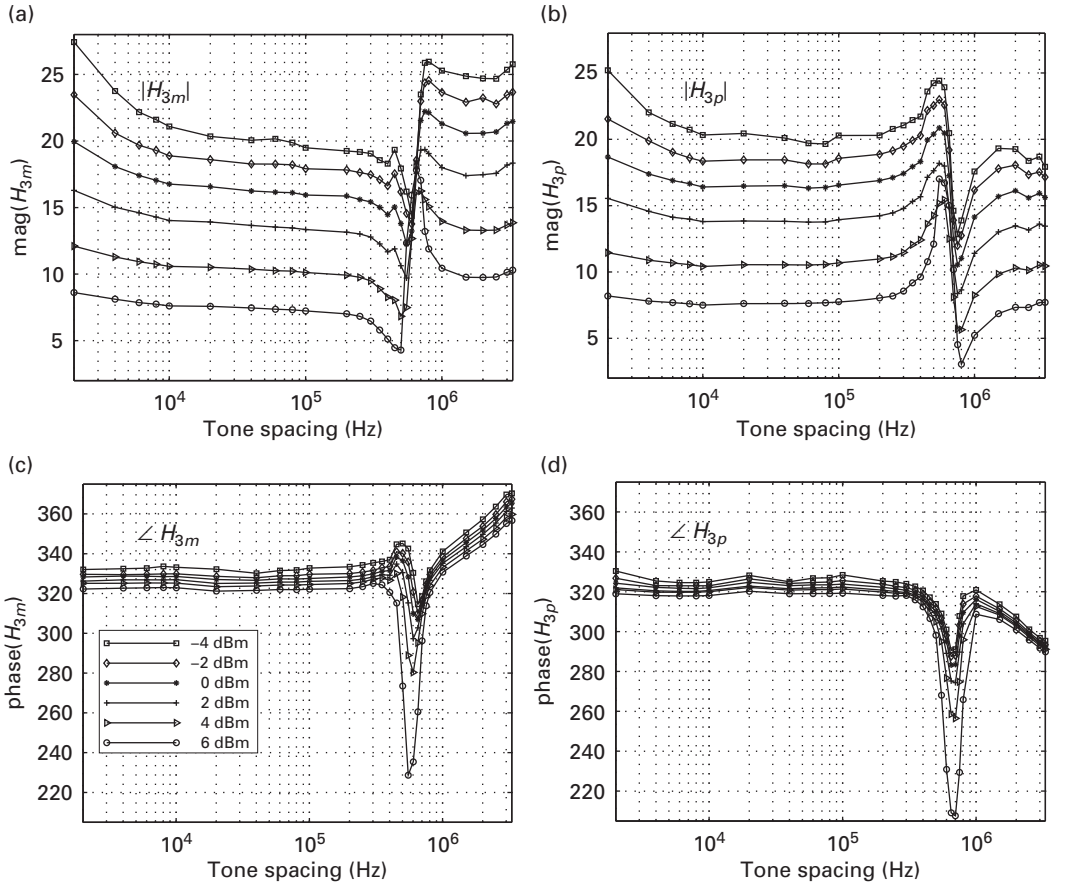


Figure 6.17 Comparison of amplitude and phase of H_{3m} and H_{3p} versus the tone spacing ω_m for different power levels (−4 to 6 dBm). (From [11] with permission, ©2005 IEEE.)

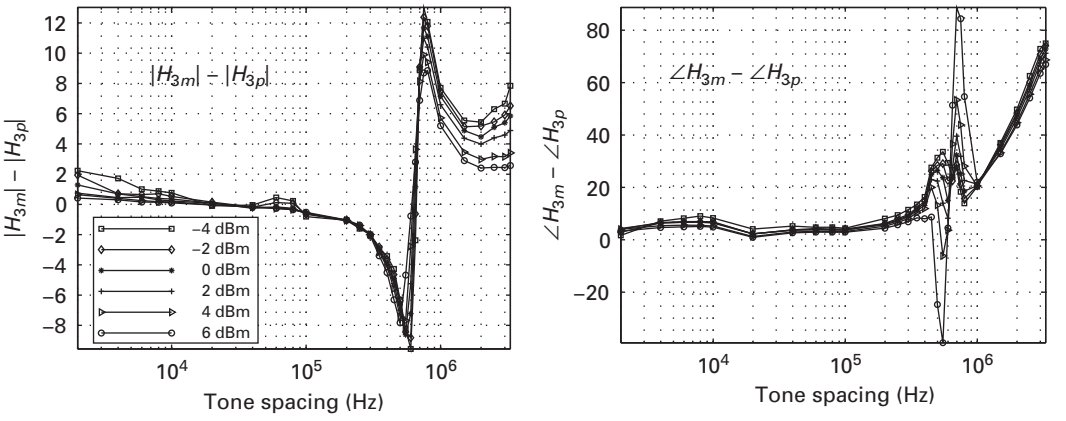


Figure 6.18 Amplitude (a) and phase (b) difference between H_{3m} and H_{3p} versus $\omega_m = \omega_2 - \omega_1$ reveals a strong *differential memory* effect above 1 MHz. (From [11] with permission, ©2005 IEEE.)

of tone spacing ω_m reveals the presence of memory effects (frequency-dependent non-linearity) in the PA. Above 0.3 MHz the *differences* in amplitude and phase plotted in Figure 6.18 increase rapidly with tone spacing. This is referred to as *differential memory*. These generalized Volterra coefficients indicate that a QML model will not be optimal for this PA for bandwidths above 0.3 MHz [11]. Note that the power dependence originates from the contribution of higher-order nonlinearities (fifth, seventh, ... orders) in the PA. A higher-order extension of the two-band model is given in the next section.

6.4.2 Extension to higher-order nonlinearities

We are again considering a power amplifier with a two-tone excitation $a_1(\omega_1)$ and $a_1(\omega_2)$ injected at the input port 1. The resulting seventh-order intermodulation terms of the output $b_{\text{out}}(\omega)$ at port 2 can be described with a Volterra system as follows [13]:

$$
\begin{aligned}
b_{\text{out}}\left(\frac{p+1}{2}\omega_1 - \frac{p-1}{2}\omega_2\right) &= f_{-p} \cdot a_1^{(p+1)/2}(\omega_1) a_1^{*(p-1)/2}(\omega_2), \\
&\cdots \\
b_{\text{out}}(4\omega_1 - 3\omega_2) &= f_{-7} \cdot a_1^4(\omega_1) a_1^{*3}(\omega_2), \\
b_{\text{out}}(3\omega_1 - 2\omega_2) &= f_{-5} \cdot a_1^3(\omega_1) a_1^{*2}(\omega_2), \qquad (6.54) \\
b_{\text{out}}(2\omega_1 - \omega_2) &= f_{-3} \cdot a_1^2(\omega_1) a_1^{*}(\omega_2), \qquad (6.55) \\
b_{\text{out}}(\omega_1) &= f_{-1} \cdot a_1(\omega_1), \qquad (6.56) \\
b_{\text{out}}(\omega_2) &= f_1 \cdot a_1(\omega_2), \qquad (6.57) \\
b_{\text{out}}(2\omega_2 - \omega_1) &= f_3 \cdot a_1^2(\omega_2) a_1^{*}(\omega_1), \qquad (6.58) \\
b_{\text{out}}(3\omega_2 - 2\omega_1) &= f_5 \cdot a_1^3(\omega_2) a_1^{*2}(\omega_1), \\
b_{\text{out}}(4\omega_2 - 3\omega_1) &= f_7 \cdot a_1^4(\omega_2) a_3^{*3}(\omega_1), \\
&\cdots \\
b_{\text{out}}\left(\frac{p+1}{2}\omega_2 - \frac{p-1}{2}\omega_1\right) &= f_p \cdot a_1^{(p+1)/2}(\omega_2) a_1^{*(p-1)/2}(\omega_1).
\end{aligned}
$$

where the Volterra functions f_{-p} and f_p can be represented as

$$
f_{-p}(|a_1(\omega_1)|^2, |a_1(\omega_2)|^2, \omega_1, \omega_2) = \frac{b_{\text{out}}\left([(p+1)/2]\omega_1 - [(p-1)/2]\omega_2\right)}{a_1^{(p+1)/2}(\omega_1) a_1^{*(p-1)/2}(\omega_2)}, \qquad (6.59)
$$

$$
f_{p}(|a_1(\omega_1)|^2, |a_1(\omega_2)|^2, \omega_1, \omega_2) = \frac{b_{\text{out}}\left([(p+1)/2]\omega_2 - [(p-1)/2]\omega_1\right)}{a_1^{*(p-1)/2}(\omega_1) a_1^{(p+1)/2}(\omega_2)}. \qquad (6.60)
$$

Note that this Volterra expansion is not valid up to infinite order (p) since, for increasing order, the intermodulation terms of the fundamental will eventually overlap with those of the higher harmonics. Let us find the maximum nonlinearity order $O_{\max}$, up to which the Volterra equations (6.54)–(6.58) provide the only contribution occuring at these frequencies. When higher-order nonlinearities are considered for the two-tone excitation at ω_0 and $\omega_0 + \Delta\omega$, the frequencies of the adjacent intermodulation products are of the form

$$\begin{aligned} f_{m,n} &= m\omega_0 + n(\omega_0 + \Delta\omega) \\ &= (m+n)\omega_0 + n\,\Delta\omega, \end{aligned} \tag{6.61}$$

with $m + n = 1$ for the fundamental band and $m + n = 2$ for the second harmonic. Let us consider a nonlinear system of order O such that we have $|n| + |m| \leq O$. The intermodulation frequencies between the adjacent fundamental and second-harmonic bands will not overlap when the highest-order upper-sideband intermodulation term for the fundamental is lower than the highest-order lower-sideband intermodulation term for the second harmonic:

$$\omega_0 + \frac{1}{2}(O+1)\Delta\omega < 2\omega_0 - \frac{1}{2}(O-2)\Delta\omega.$$

$O_{\max}$ is then the highest even-order integer satisfying this inequality. The maximum order $O_{\max}$ of the Volterra system before the sideband tones of the adjacent harmonic start overlapping is thus given by

$$O_{\max} < \frac{1}{2} + \frac{\omega_0}{\Delta\omega}. \tag{6.62}$$

The maximum order $O_{\max}$ obtained from Equation (6.62) is 10 and 100 for a communication system with 10% and 1% fractional bandwidth, respectively. The two-band Volterra expansion is seen to hold in practice for a relatively wide bandwidth since intermodulation terms of higher than tenth order are usually negligible in PAs intended for communication systems.

6.4.3 Modulated two-band model

We shall now discuss the extension of this Volterra model to the case of narrow-bandwidth modulated signals. Again we focus on the fundamental frequency. When the baseband signal (I, Q) consists of two narrow sub-bands, say a lower sideband (LSB) $(I_{\mathrm{L}}, Q_{\mathrm{L}})$ and upper sideband (USB) $(I_{\mathrm{U}}, Q_{\mathrm{U}})$ with center frequencies that are far enough apart, it becomes advantageous to model the baseband signal as consisting of two distinct baseband signals rather than to keep increasing the number of delays in the time-selective model or memory polynomials to account for the wide modulation bandwidth memory effects associated with their respective memory effects. The easiest way to perform this two-band filtering is to partition the input I and Q in terms of their USB and LSB components as follows:

$$\begin{aligned} I(t) &= I_{\mathrm{U}}(t) + I_{\mathrm{L}}(t), \\ Q(t) &= Q_{\mathrm{U}}(t) + Q_{\mathrm{L}}(t), \\ Q_{\mathrm{U}} &= \widehat{I_{\mathrm{U}}}, \\ Q_{\mathrm{L}} &= -\widehat{I_{\mathrm{L}}}. \end{aligned}$$

Note that we use the notation $\widehat{I(t)} = \mathcal{H}[I(t)]$ to denote the Hilbert transform of $I(t)$. One can readily verify that I_{U}, I_{L}, $\widehat{I_{\mathrm{U}}}$, and $\widehat{I_{\mathrm{L}}}$ are obtained from I and Q and their Hilbert transforms by using

$$I_{\mathrm{L}} = \frac{I + \widehat{Q}}{2}, \qquad \widehat{I_{\mathrm{L}}} = \frac{\widehat{I} - Q}{2}, \qquad I_{\mathrm{U}} = \frac{I - \widehat{Q}}{2}, \qquad \widehat{I_{\mathrm{U}}} = \frac{\widehat{I} + Q}{2}.$$

Consider a nonlinearity of order p with p odd and positive ($p = |p|$). Given two frequencies f_1 and f_2 with $\Delta f = f_2 - f_1 > 0$, this odd nonlinearity will generate intermodulation products around the fundamental frequencies $f_0 = (f_2 + f_1)/2$ at the frequencies

$$\begin{aligned} f_p &= (n_p + 1) f_2 - n_p f_1 = f_2 + n_p \Delta f = f_0 + \left(n_p + \frac{1}{2}\right) \Delta f = f_0 + \Delta f_p, \\ f_{-p} &= (n_p + 1) f_1 - n_p f_2 = f_1 - n_p \Delta f = f_0 - \left(n_p + \frac{1}{2}\right) \Delta f = f_0 + \Delta f_{-p}, \end{aligned} \tag{6.63}$$

with

$$n_p = \frac{|p| - 1}{2} \qquad \text{and} \qquad \Delta f_{\pm p} = \pm \left(n_p + \frac{1}{2}\right) \Delta f.$$

Let us assume the two tones are amplitude- and phase-modulated such that we can write

$$\begin{aligned} I_{\mathrm{L}}(t) &= E_{\mathrm{L}}(t) \cos\left(\tfrac{1}{2} \Delta\omega\, t + \phi_{\mathrm{L}}(t)\right), \\ Q_{\mathrm{L}}(t) &= -E_{\mathrm{L}}(t) \sin\left(\tfrac{1}{2} \Delta\omega\, t + \phi_{\mathrm{L}}(t)\right), \\ I_{\mathrm{U}}(t) &= E_{\mathrm{U}}(t) \cos\left(\tfrac{1}{2} \Delta\omega\, t + \phi_{\mathrm{U}}(t)\right), \\ Q_{\mathrm{U}}(t) &= E_{\mathrm{U}}(t) \sin\left(\tfrac{1}{2} \Delta\omega\, t + \phi_{\mathrm{U}}(t)\right), \end{aligned} \tag{6.64}$$

with $\Delta\omega = 2\pi \Delta f$. The intermodulation terms generated will be of the form

$$\begin{aligned} I_p &= E_{\mathrm{L}}^{n_p} E_{\mathrm{U}}^{n_p+1} \cos\left[\Delta\omega_p + (n_p + 1)\phi_{\mathrm{U}} - n_p \phi_{\mathrm{L}}\right], \\ Q_p &= E_{\mathrm{L}}^{n_p} E_{\mathrm{U}}^{n_p+1} \sin\left[\Delta\omega_p + (n_p + 1)\phi_{\mathrm{U}} - n_p \phi_{\mathrm{L}}\right], \\ I_{-p} &= E_{\mathrm{U}}^{n_p} E_{\mathrm{L}}^{n_p+1} \cos\left[\Delta\omega_p + (n_p + 1)\phi_{\mathrm{L}} - n_p \phi_{\mathrm{U}}\right], \\ Q_{-p} &= E_{\mathrm{U}}^{n_p} E_{\mathrm{L}}^{n_p+1} \sin\left[\Delta\omega_p + (n_p + 1)\phi_{\mathrm{L}} - n_p \phi_{\mathrm{U}}\right]. \end{aligned} \tag{6.65}$$

It is quite impractical (costly and less accurate) in a digital signal processing implementation to calculate the instantaneous phases $\phi_{\mathrm{U}}(t)$ and $\phi_{\mathrm{L}}(t)$. An alternative approach is to calculate these intermodulation terms directly from (I_{L}, Q_{L}) and (I_{U}, Q_{U}). The following identities can be established:

$$\begin{aligned} I_1 &= I_{\mathrm{U}}, \\ Q_1 &= Q_{\mathrm{U}}, \\ I_{-1} &= I_{\mathrm{L}}, \\ Q_{-1} &= Q_{\mathrm{L}}, \\ I_3 &= I_{\mathrm{U}}^2 I_{\mathrm{L}} + 2 I_{\mathrm{U}} Q_{\mathrm{U}} Q_{\mathrm{L}} - Q_{\mathrm{U}}^2 I_{\mathrm{L}}, \\ Q_3 &= -I_{\mathrm{U}}^2 Q_{\mathrm{L}} + 2 I_{\mathrm{U}} Q_{\mathrm{U}} I_{\mathrm{L}} + Q_{\mathrm{U}}^2 Q_{\mathrm{L}}, \\ I_{-3} &= I_{\mathrm{U}} I_{\mathrm{L}}^2 - I_{\mathrm{U}} Q_{\mathrm{L}}^2 + 2 Q_{\mathrm{U}} I_{\mathrm{L}} Q_{\mathrm{L}}, \end{aligned}$$

$$
\begin{aligned}
Q_{-3} &= 2I_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}} - Q_{\mathrm{U}}I_{\mathrm{L}}^2 + Q_{\mathrm{U}}Q_{\mathrm{L}}^2,\\
I_5 &= I_{\mathrm{U}}^3I_{\mathrm{L}}^2 - I_{\mathrm{U}}^3Q_{\mathrm{L}}^2 + 6I_{\mathrm{U}}^2Q_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}} - 3I_{\mathrm{U}}Q_{\mathrm{U}}^2I_{\mathrm{L}}^2 + 3I_{\mathrm{U}}Q_{\mathrm{U}}^2Q_{\mathrm{L}}^2 - 2Q_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}},\\
Q_5 &= -2I_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}} + 3I_{\mathrm{U}}^2Q_{\mathrm{U}}I_{\mathrm{L}}^2 - 3I_{\mathrm{U}}^2Q_{\mathrm{U}}Q_{\mathrm{L}}^2 + 6I_{\mathrm{U}}Q_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}} - Q_{\mathrm{U}}^3I_{\mathrm{L}}^2 + Q_{\mathrm{U}}^3Q_{\mathrm{L}}^2,\\
I_{-5} &= I_{\mathrm{U}}^2I_{\mathrm{L}}^3 - 3I_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}}^2 + 6I_{\mathrm{U}}Q_{\mathrm{U}}I_{\mathrm{L}}^2Q_{\mathrm{L}} - 2I_{\mathrm{U}}Q_{\mathrm{U}}Q_{\mathrm{L}}^3 - Q_{\mathrm{U}}^2I_{\mathrm{L}}^3 + 3Q_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}}^2,\\
Q_{-5} &= 3I_{\mathrm{U}}^2I_{\mathrm{L}}^2Q_{\mathrm{L}} - I_{\mathrm{U}}^2Q_{\mathrm{L}}^3 - 2I_{\mathrm{U}}Q_{\mathrm{U}}I_{\mathrm{L}}^3 + 6I_{\mathrm{U}}Q_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}}^2 - 3Q_{\mathrm{U}}^2I_{\mathrm{L}}^2Q_{\mathrm{L}} + Q_{\mathrm{U}}^2Q_{\mathrm{L}}^3,\\
I_7 &= I_{\mathrm{U}}^4I_{\mathrm{L}}^3 - 3I_{\mathrm{U}}^4I_{\mathrm{L}}Q_{\mathrm{L}}^2 + 12I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}^2Q_{\mathrm{L}} - 4I_{\mathrm{U}}^3Q_{\mathrm{U}}Q_{\mathrm{L}}^3 - 6I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}^3\\
&\quad + 18I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}}^2 - 12I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}^2Q_{\mathrm{L}} + 4I_{\mathrm{U}}Q_{\mathrm{U}}^3Q_{\mathrm{L}}^3 + Q_{\mathrm{U}}^4I_{\mathrm{L}}^3 - 3Q_{\mathrm{U}}^4I_{\mathrm{L}}Q_{\mathrm{L}}^2,\\
Q_7 &= -3I_{\mathrm{U}}^4I_{\mathrm{L}}^2Q_{\mathrm{L}} + I_{\mathrm{U}}^4Q_{\mathrm{L}}^3 + 4I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}^3 - 12I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}}^2 + 18I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}^2Q_{\mathrm{L}}\\
&\quad - 6I_{\mathrm{U}}^2Q_{\mathrm{U}}^2Q_{\mathrm{L}}^3 - 4I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}^3 + 12I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}}^2 - 3Q_{\mathrm{U}}^4I_{\mathrm{L}}^2Q_{\mathrm{L}} + Q_{\mathrm{U}}^4Q_{\mathrm{L}}^3,\\
I_{-7} &= I_{\mathrm{U}}^3I_{\mathrm{L}}^4 - 6I_{\mathrm{U}}^3I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 + I_{\mathrm{U}}^3Q_{\mathrm{L}}^4 + 12I_{\mathrm{U}}^2Q_{\mathrm{U}}I_{\mathrm{L}}^3Q_{\mathrm{L}} - 12I_{\mathrm{U}}^2Q_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}}^3 - 3I_{\mathrm{U}}Q_{\mathrm{U}}^2I_{\mathrm{L}}^4\\
&\quad + 18I_{\mathrm{U}}Q_{\mathrm{U}}^2I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 - 3I_{\mathrm{U}}Q_{\mathrm{U}}^2Q_{\mathrm{L}}^4 - 4Q_{\mathrm{U}}^3I_{\mathrm{L}}^3Q_{\mathrm{L}} + 4Q_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}}^3,\\
Q_{-7} &= 4I_{\mathrm{U}}^3I_{\mathrm{L}}^3Q_{\mathrm{L}} - 4I_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}}^3 - 3I_{\mathrm{U}}^2Q_{\mathrm{U}}I_{\mathrm{L}}^4 + 18I_{\mathrm{U}}^2Q_{\mathrm{U}}I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 - 3I_{\mathrm{U}}^2Q_{\mathrm{U}}Q_{\mathrm{L}}^4\\
&\quad - 12I_{\mathrm{U}}Q_{\mathrm{U}}^2I_{\mathrm{L}}^3Q_{\mathrm{L}} + 12I_{\mathrm{U}}Q_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}}^3 + Q_{\mathrm{U}}^3I_{\mathrm{L}}^4 - 6Q_{\mathrm{U}}^3I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 + Q_{\mathrm{U}}^3Q_{\mathrm{L}}^4,\\
I_9 &= I_{\mathrm{U}}^5I_{\mathrm{L}}^4 - 6I_{\mathrm{U}}^5I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 + I_{\mathrm{U}}^5Q_{\mathrm{L}}^4 + 20I_{\mathrm{U}}^4Q_{\mathrm{U}}I_{\mathrm{L}}^3Q_{\mathrm{L}} - 20I_{\mathrm{U}}^4Q_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}}^3 - 10I_{\mathrm{U}}^3Q_{\mathrm{U}}^2I_{\mathrm{L}}^4\\
&\quad + 60I_{\mathrm{U}}^3Q_{\mathrm{U}}^2I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 - 10I_{\mathrm{U}}^3Q_{\mathrm{U}}^2Q_{\mathrm{L}}^4 - 40I_{\mathrm{U}}^2Q_{\mathrm{U}}^3I_{\mathrm{L}}^3Q_{\mathrm{L}} + 40I_{\mathrm{U}}^2Q_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}}^3,\\
&\quad + 5I_{\mathrm{U}}Q_{\mathrm{U}}^4I_{\mathrm{L}}^4 - 30I_{\mathrm{U}}Q_{\mathrm{U}}^4I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 + 5I_{\mathrm{U}}Q_{\mathrm{U}}^4Q_{\mathrm{L}}^4 + 4Q_{\mathrm{U}}^5I_{\mathrm{L}}^3Q_{\mathrm{L}} - 4Q_{\mathrm{U}}^5I_{\mathrm{L}}Q_{\mathrm{L}}^3,\\
Q_9 &= -4I_{\mathrm{U}}^5I_{\mathrm{L}}^3Q_{\mathrm{L}} + 4I_{\mathrm{U}}^5I_{\mathrm{L}}Q_{\mathrm{L}}^3 + 5I_{\mathrm{U}}^4Q_{\mathrm{U}}I_{\mathrm{L}}^4 - 30I_{\mathrm{U}}^4Q_{\mathrm{U}}I_{\mathrm{L}}^2Q_{\mathrm{L}}^2 + 5I_{\mathrm{U}}^4Q_{\mathrm{U}}Q_{\mathrm{L}}^4\\
&\quad + 40I_{\mathrm{U}}^3Q_{\mathrm{U}}^2I_{\mathrm{L}}^3Q_{\mathrm{L}} - 40I_{\mathrm{U}}^3Q_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}}^3 - 10I_{\mathrm{U}}^2Q_{\mathrm{U}}^3I_{\mathrm{L}}^4 + 60I_{\mathrm{U}}^2Q_{\mathrm{U}}^3I_{\mathrm{L}}^2Q_{\mathrm{L}}^2\\
&\quad - 10I_{\mathrm{U}}^2Q_{\mathrm{U}}^3Q_{\mathrm{L}}^4 - 20I_{\mathrm{U}}Q_{\mathrm{U}}^4I_{\mathrm{L}}^3Q_{\mathrm{L}} + 20I_{\mathrm{U}}Q_{\mathrm{U}}^4I_{\mathrm{L}}Q_{\mathrm{L}}^3 + Q_{\mathrm{U}}^5I_{\mathrm{L}}^4 - 6Q_{\mathrm{U}}^5I_{\mathrm{L}}^2Q_{\mathrm{L}}^2\\
&\quad + Q_{\mathrm{U}}^5Q_{\mathrm{L}}^4,\\
I_{-9} &= I_{\mathrm{U}}^4I_{\mathrm{L}}^5 - 10I_{\mathrm{U}}^4I_{\mathrm{L}}^3Q_{\mathrm{L}}^2 + 5I_{\mathrm{U}}^4I_{\mathrm{L}}Q_{\mathrm{L}}^4 + 20I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}^4Q_{\mathrm{L}} - 40I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}^2Q_{\mathrm{L}}^3\\
&\quad + 4I_{\mathrm{U}}^3Q_{\mathrm{U}}Q_{\mathrm{L}}^5 - 6I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}^5 + 60I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}^3Q_{\mathrm{L}}^2 - 30I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}Q_{\mathrm{L}}^4\\
&\quad - 20I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}^4Q_{\mathrm{L}} + 40I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}^2Q_{\mathrm{L}}^3 - 4I_{\mathrm{U}}Q_{\mathrm{U}}^3Q_{\mathrm{L}}^5 + Q_{\mathrm{U}}^4I_{\mathrm{L}}^5 - 10Q_{\mathrm{U}}^4I_{\mathrm{L}}^3Q_{\mathrm{L}}^2\\
&\quad + 5Q_{\mathrm{U}}^4I_{\mathrm{L}}Q_{\mathrm{L}}^4,\\
Q_{-9} &= 5I_{\mathrm{U}}^4I_{\mathrm{L}}^4Q_{\mathrm{L}} - 10I_{\mathrm{U}}^4I_{\mathrm{L}}^2Q_{\mathrm{L}}^3 + I_{\mathrm{U}}^4Q_{\mathrm{L}}^5 - 4I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}^5 + 40I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}^3Q_{\mathrm{L}}^2\\
&\quad - 20I_{\mathrm{U}}^3Q_{\mathrm{U}}I_{\mathrm{L}}Q_{\mathrm{L}}^4 - 30I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}^4Q_{\mathrm{L}} + 60I_{\mathrm{U}}^2Q_{\mathrm{U}}^2I_{\mathrm{L}}^2Q_{\mathrm{L}}^3 - 6I_{\mathrm{U}}^2Q_{\mathrm{U}}^2Q_{\mathrm{L}}^5\\
&\quad + 4I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}^5 - 40I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}^3Q_{\mathrm{L}}^2 + 20I_{\mathrm{U}}Q_{\mathrm{U}}^3I_{\mathrm{L}}Q_{\mathrm{L}}^4 + 5Q_{\mathrm{U}}^4I_{\mathrm{L}}^4Q_{\mathrm{L}} - 10Q_{\mathrm{U}}^4I_{\mathrm{L}}^2Q_{\mathrm{L}}^3\\
&\quad + Q_{\mathrm{U}}^4Q_{\mathrm{L}}^5.
\end{aligned}
$$

Similarly to the harmonic terms in the Chaillot expansion introduced in the previous sections, these intermodulation terms will be nonlinearly scaled and phase-shifted as we account for higher-order nonlinearities.

The new intermodulation terms I'_p and Q'_p become then

$$\begin{bmatrix} I'_p(t) \\ Q'_p(t) \end{bmatrix} = \begin{bmatrix} g_p\left(E_{\mathrm{L}}^2, E_{\mathrm{U}}^2\right) & -h_p\left(E_{\mathrm{L}}^2, E_{\mathrm{U}}^2\right) \\ h_p\left(E_{\mathrm{L}}^2, E_{\mathrm{U}}^2\right) & g_p\left(E_{\mathrm{L}}^2, E_{\mathrm{U}}^2\right) \end{bmatrix} \times \begin{bmatrix} I_p(t) \\ Q_p(t) \end{bmatrix}. \tag{6.66}$$

The output of the amplifier in the fundamental band for a two-band excitation is then given by

$$\begin{aligned} I_{\mathrm{out}}(t) &= \sum_{p=-P}^{P} I'_p(t), \\ Q_{\mathrm{out}}(t) &= \sum_{p=-P}^{P} Q'_p(t), \\ b_{\mathrm{out}}^{\mathrm{QML}}(t) &= \sum_{p=-P}^{P} [I_{\mathrm{out}}(t)\cos(\omega_0 t) - Q_{\mathrm{out}}(t)\sin(\omega_0 t)]. \end{aligned} \tag{6.67}$$

Note that now the functions g_p and h_p are functions of both the time-varying envelopes $E_{\mathrm{U}}(t)$ and $E_{\mathrm{L}}(t)$ since both of these terms bring high-order DC corrections. Again the functions g_p and h_p are conceptually accounting for infinite-order correction. However, this is rigorous only in the limit of constant upper- and lower-sideband envelopes. Indeed, the memory effects associated with the time-varying $(I_{\mathrm{L}}, Q_{\mathrm{L}})$ and $(I_{\mathrm{U}}, Q_{\mathrm{U}})$ components in each of the intermodulation sub-bands have not been accounted for in non-zero modulation bandwidths. The model is therefore only piecewise quasi-memoryless and only the AM/AM and AM/PM effects have been accounted for in each of the intermodulation bands. Again memory effects can then be accounted for using memory polynomials as was demonstrated for the single-band case.

Note that the intermodulation band of index p exhibits a bandwidth of $(2n_p + 1)B$, with B the individual bandwidth of the desired bands centered at f_1 and f_2. Since the bandwith of each intermodulation band increases with the intermodulation index p and the data bandwidth B, the various intermodulation bands will start overlapping for large enough index p and wide enough bandwidth B. This points to the interplay between bandwidth and nonlinearity in nonlinear devices. To benefit from the frequency selectivity in modeling and linearization, the two-band model is therefore most beneficially applied to two-band communication PA systems rather than two-carrier PA systems owing to their larger frequency separation.

6.5 Appendix: Volterra series expansion for a four-tone excitation

The reflected waves $b(\omega_0)$, $b(2\omega_0)$, and $b(3\omega_0)$ for four incident tones can be represented using the following Volterra series expansion:

$$\begin{aligned} b(\omega_0) = {} & a(\omega_0) \cdot V_{1,1}(\mathbf{x}_4) + a^*(\omega_0)a(2\omega_0) \cdot V_{1,2}(\mathbf{x}_4) \\ & + a^*(2\omega_0)a(3\omega_0) \cdot V_{1,3}(\mathbf{x}_4) + a^{*2}(\omega_0)a(3\omega_0) \cdot V_{1,4}(\mathbf{x}_4) \end{aligned}$$

$$
\begin{aligned}
&+ a^*(3\omega_0)a^2(2\omega_0)\cdot V_{1,5}(\mathbf{x}_4) + a^*(3\omega_0)a(4\omega_0)\cdot V_{1,6}(\mathbf{x}_4)\\
&+ a^{*2}(4\omega_0)a^3(3\omega_0)\cdot V_{1,7}(\mathbf{x}_4) + a^{*3}(\omega_0)a(4\omega_0)\cdot V_{1,8}(\mathbf{x}_4)\\
&+ a^*(2\omega_0)a^*(\omega_0)a(4\omega_0)\cdot V_{1,9}(\mathbf{x}_4) + a^*(4\omega_0)a(2\omega_0)a(3\omega_0)\cdot V_{1,10}(\mathbf{x}_4)\\
&+ a^*(4\omega_0)a^*(\omega_0)a^2(3\omega_0)\cdot V_{1,11}(\mathbf{x}_4),
\end{aligned}
\tag{6.68}
$$

$$
\begin{aligned}
b(2\omega_0) = {} & a^2(\omega_0)\cdot V_{2,1}(\mathbf{x}_4) + a(2\omega_0)\cdot V_{2,2}(\mathbf{x}_4)\\
&+ a^*(2\omega_0)a(\omega_0)a(3\omega_0)\cdot V_{2,3}(\mathbf{x}_4) + a^*(\omega_0)a(3\omega_0)\cdot V_{2,4}(\mathbf{x}_4)\\
&+ a^{*2}(2\omega_0)a^2(3\omega_0)\cdot V_{2,5}(\mathbf{x}_4) + a^*(2\omega_0)a(4\omega_0)\cdot V_{2,6}(\mathbf{x}_4)\\
&+ a^*(4\omega_0)a^2(3\omega_0)\cdot V_{2,7}(\mathbf{x}_4) + a^{*2}(\omega_0)a(4\omega_0)\cdot V_{2,8}(\mathbf{x}_4)\\
&+ a^{*2}(3\omega_0)a^2(4\omega_0)\cdot V_{2,9}(\mathbf{x}_4) + a^*(3\omega_0)a(\omega_0)a(4\omega_0)\cdot V_{2,10}(\mathbf{x}_4),
\end{aligned}
\tag{6.69}
$$

$$
\begin{aligned}
b(3\omega_0) = {} & a^3(\omega_0)\cdot V_{3,1}(\mathbf{x}_4) + a(\omega_0)a(2\omega_0)\cdot V_{3,2}(\mathbf{x}_4)\\
&+ a^*(\omega_0)a^2(2\omega_0)\cdot V_{3,3}(\mathbf{x}_4) + a(3\omega_0)\cdot V_{3,4}(\mathbf{x}_4)\\
&+ a^*(3\omega_0)a^3(2\omega_0)\cdot V_{3,5}(\mathbf{x}_4) + a^*(\omega_0)a(4\omega_0)\cdot V_{3,6}(\mathbf{x}_4)\\
&+ a^{*3}(3\omega_0)a^3(4\omega_0)\cdot V_{3,7}(\mathbf{x}_4) + a^*(3\omega_0)a^*(2\omega_0)a^2(4\omega_0)\cdot V_{3,8}(\mathbf{x}_4)\\
&+ a^*(3\omega_0)a^2(\omega_0)a(4\omega_0)\cdot V_{3,9}(\mathbf{x}_4) + a^*(3\omega_0)a(2\omega_0)a(4\omega_0)\cdot V_{3,10}(\mathbf{x}_4)\\
&+ a^*(2\omega_0)a(\omega_0)a(4\omega_0)\cdot V_{3,11}(\mathbf{x}_4) + a^{*2}(3\omega_0)a(\omega_0)a^2(4\omega_0)\cdot V_{3,12}(\mathbf{x}_4),
\end{aligned}
\tag{6.70}
$$

where each of the 32 functions $V_{n,i}(\mathbf{x}_4, n\omega_0)$ in (6.68)–(6.70) was found to be functionally dependent on 34 DC terms $x_{4,i}$:

$$
V_{n,i}(\mathbf{x}_4) = V_{n,i}(x_{4,1}, x_{4,2}, \ldots, x_{4,34}). \tag{6.71}
$$

These DC terms are given by

$$
\begin{aligned}
x_{4,1} &= a^*(\omega_0)a(\omega_0), & x_{4,2} &= a^*(2\omega_0)a(2\omega_0),\\
x_{4,3} &= a^*(2\omega_0)a^2(\omega_0), & x_{4,4} &= x_{4,3}^*,\\
x_{4,5} &= a^*(3\omega_0)a(3\omega_0), & &\\
x_{4,6} &= a^*(3\omega_0)a(\omega_0)a(2\omega_0), & x_{4,7} &= x_{4,6}^*,\\
x_{4,8} &= a^*(3\omega_0)a^3(\omega_0), & x_{4,9} &= x_{4,8}^*,\\
x_{4,10} &= a^*(3\omega_0)a^*(\omega_0)a^2(2\omega_0), & x_{4,11} &= x_{4,10}^*,\\
x_{4,12} &= a^{*2}(3\omega_0)a^3(2\omega_0), & x_{4,13} &= x_{4,12}^*,\\
x_{4,14} &= a^*(4\omega_0)a(4\omega_0), & &\\
x_{4,15} &= a^{*2}(4\omega_0)a^*(\omega_0)a^3(3\omega_0), & x_{4,17} &= a^*(4\omega_0)a(\omega_0)a(3\omega_0),\\
x_{4,19} &= a^*(4\omega_0)a^2(2\omega_0), & x_{4,21} &= a^*(4\omega_0)a^*(2\omega_0)a^2(3\omega_0),\\
x_{4,23} &= a^*(4\omega_0)a^4(\omega_0), & x_{4,25} &= a^*(4\omega_0)a^{*2}(\omega_0)a(2\omega_0),\\
x_{4,27} &= a^{*3}(4\omega_0)a^4(3\omega_0), & x_{4,29} &= a^*(4\omega_0)a^{*2}(\omega_0)a^2(3\omega_0),\\
x_{4,31} &= a^{*2}(4\omega_0)a(2\omega)a^2(3\omega_0), & x_{4,33} &= a^*(4\omega_0)a^*(\omega_0)a(2\omega_0)a(3\omega_0),\\
x_{4,i} &= x_{4,i-1}^*, \quad \text{for } 16 < i \text{ (even)} < 34. & &
\end{aligned}
$$

References

[1] D. Schreurs, M. O'Drama, A. Goacher, and M. Gadringer, *RF Power Amplifier Behavioral Modeling*, Cambridge University Press, 2008.

[2] J. C. Peyton Jones and S. A Billings, "Describing functions, Volterra series, and the analysis of non-linear systems in the frequency domain," *International Journal of Control*, Vol. 53, No. 4, pp. 871–887, April 1991.

[3] A. A. M. Saleh, "Matrix analysis of mildly nonlinear, multiple-input, multiple-output systems with memory," *The Bell System Technical Journal*, Vol. 61, No. 9, pp. 2221–2243, Nov. 1982.

[4] M. Schetzen "Nonlinear system modeling based on the Wiener theory," *Proceedings of the IEEE*, Vol. 69, No. 12, pp. 1557–1573, Dec. 1981.

[5] J. J. Bussgang, L. Ehrman, and J. W. Graham, "Analysis of nonlinear systems with multiple inputs," *Proceedings of the IEEE*, Vol. 62, No. 8, pp. 1088–1119, Dec. 1974.

[6] J. C. Predro and N. B. Carvalho, *Intermodulation Distortion in Microwave and Wireless Circuits*, Artech House, 2003.

[7] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters*, Vol. 37, No. 23, pp. 1417–1418, Dec. 2001.

[8] L. Ding, G. T. Xhou, D. R. Morgan, Z. Ma, J. S. Kenny, J. Kim, and C. R. Giardina, "A robust digital baseband predistorter constructed using memory polynomials," *IEEE Journal of Communication*, Vol. 52, No. 1, pp. 159–165, Jan. 2004.

[9] D. E. Root, J. Verspecht, D. Sharrit, J. Wood, and A. Cognata, "Broad-band polyharmonic distortion (PHD) behavioral models from fast automated simulations and large-signal scattering functions," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 53, No. 11, pp. 3656–3664, Nov. 2005.

[10] J. S. Kenney, W. Woo, L. Ding, R. Raich, H. Ku and G. T. Zhou, "The impact of memory effects on predistortion linearization of RF power amplifiers," *Proceedings of the 8th International Symposium on Microwave and Optical Techniques*, Montreal, pp. 189–193, 2001.

[11] S. K. Myoung, D. Chaillot, P. Roblin, W. Dai, and S. J. Doo, "Volterra characterization and predistortion linearization of multi-carrier power amplifiers," in *64th ARFTG Conference Digest,* Orlando, FL, 2004.

[12] J. Verspecht, F. Verspecht, and M. Vanden Bossche, "Network analysis beyond S-parameters," in *56th ARFTG Conference Digest*, Broomfield, CO, 2000.

[13] P. Roblin, S. K. Myoung, D. Chaillot, Y. G. Kim, A. Fathimulla, J. Strahler, and S. Bibyk, "Frequency-selective predistortion linearization of RF power amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 56, No. 1, pp. 65–76, Jan. 2008.

7 Kurokawa theory of oscillator design and phase-noise theory[1]

Oscillators are essential components of radio-frequency (RF) transceivers in wireless communication systems. Together with mixers, they are used for frequency translation. Oscillators are nonlinear autonomous circuits that establish their own operating frequency and amplitude of oscillation. In this chapter we will review the nonlinear Kurokawa theory developed for the design of oscillators. The Kurokawa theory is able to account for the device nonlinearity to predict the oscillator operating point (frequency, amplitude) and establish the stability of the operating point. Experimental vector measurements of the nonlinear device line with an NVNA will then be presented. Next the Kurokawa theory will be used to derive the oscillator phase noise and amplitude noise for both white and flicker noise while accounting for amplitude and phase-noise correlation. A comparison with other theories and simulation results will be presented. Finally, the application of the Kurokawa theory to injection locking will be reviewed and a setup for simultaneous large-signal RF measurements and additive noise measurements of an injection-locked oscillator will be introduced.

7.1 Oscillator operating point

Consider an oscillator with a shunt resonator represented by the admittance circuit shown in Figure 7.1. The impedance Y_{L} is the admittance of the resonator circuit (load). Y_{IN} is the admittance of the active device which provides a nonlinear negative conductance at the resonance frequency.

The voltage across the load and the active device can be represented in terms of its harmonics by

$$v(t) = \mathrm{Re}\left\{\sum_{n=-N}^{N} V_n \exp\left[jn(\omega_0 t + \phi)\right]\right\},$$

where N is the number considered in the analysis. Note that ϕ is the phase for the first harmonic ω_0; V_n are complex numbers except for the fundamental frequency, where V_1 is real and positive and therefore equal to its amplitude $V_1 = |V_1| = A$. The current in the load is then given by

[1] Research collaboration with Inwon Suh and Jayanta Mukherjee is gratefully acknowledged.

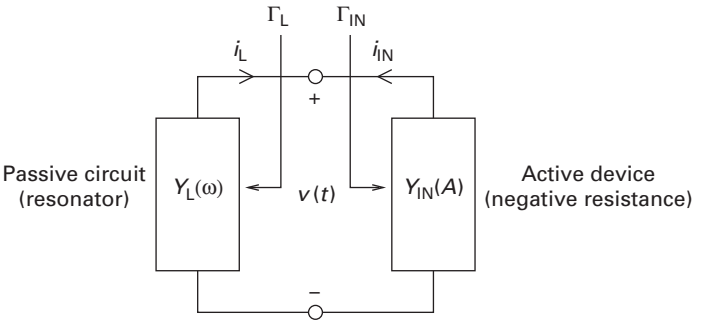


Figure 7.1 Admittance equivalent circuit for an oscillator at the fundamental frequency in steady state.

$$i_L(t) = \mathrm{Re}\left\{\sum_{n=-N}^{N} I_n \exp\left[jn(\omega_0 t + \phi)\right]\right\}$$
$$= \mathrm{Re}\left\{\sum_{n=-N}^{N} Y_L(n\omega_0) V_n \exp\left[jn(\omega_0 t + \phi)\right]\right\}.$$

The current in the active device is given by

$$i_{IN}(t) = \mathrm{Re}\left\{\sum_{n=-N}^{N} I_n \exp\left[jn(\omega_0 t + \phi)\right]\right\}$$
$$= \mathrm{Re}\left\{\sum_{n=-N}^{N} Y_{IN}(A_0, V_2, \ldots, V_N, n\omega_0) V_n \exp\left[jn(\omega_0 t + \phi)\right]\right\}.$$

Note that the admittance $Y_{IN}(A_0, V_2, \ldots, V_N, n\omega_0)$ of the nonlinear device is a function of the amplitude A_0 and the complex harmonic coefficients V_n for n larger than one.

In steady-state operation the current satisfies $i_{IN} + i_L = 0$, and it results that we have for each harmonic

$$Y_{IN}(A_0, V_2, \ldots, V_N, n\omega_0) + Y_L(n\omega_0) = 0.$$

Expressed in terms of reflection coefficients, we can rewrite this equation as

$$\Gamma_{IN}(A_0, V_2, \ldots, V_N, n\omega_0) \times \Gamma_L(n\omega_0) = 1.$$

In harmonic-balance circuit simulations, the circuit shown in Figure 7.2 is used to obtain the solution of these equations. The frequency-dependent S-matrix $\mathbf{S}(\omega_0)$ is given by

$$\mathbf{S}(\omega_0) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{S}(n\omega_0) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It results that the harmonics $n\omega$ (with $n > 1$) are terminated by the passive circuit so that the circuit operates in closed loop for them. However, an open-loop circuit

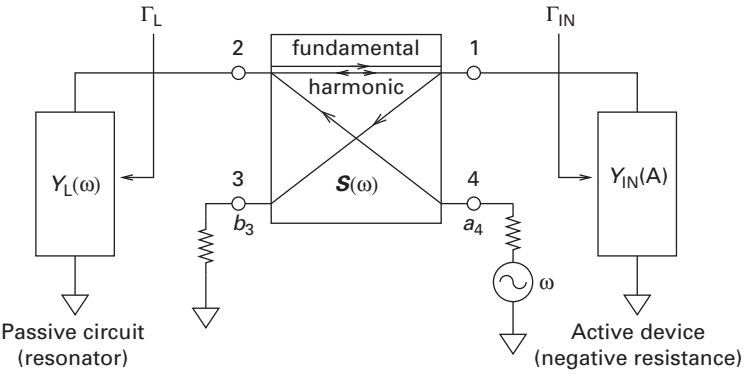


Figure 7.2 Oscillator circuit used for harmonic-balance analysis.

is generated for the fundamental frequency. The loop gain is then obtained from the measurement

$$\frac{b_3(\omega)}{a_4(\omega)} = \Gamma_{IN}[A, \Gamma_L(2\omega), \ldots, \Gamma_L(N\omega), \omega] \times \Gamma_L(\omega),$$

where the dependence of Γ_{IN} on the harmonic impedance terminations $\Gamma_L(n\omega)$ is indicated. The condition for sustained oscillation reduces then to

$$\Gamma_{IN}[A_0, \Gamma_L(2\omega_0), \ldots, \Gamma_L(N\omega_0), \omega_0] \times \Gamma_L(\omega_0) = 1.$$

The explicit dependence on the harmonic impedance terminations $\Gamma_L(n\omega)$ reminds us that the device line and the operating point can be modified by changing the harmonic impedance terminations. This invites the optimization of the oscillator performance using a load-pull, as will be considered in Section 7.3.

7.2 Kurokawa theory of oscillators

The steady-state nonlinear analysis reported in the previous section enables one to obtain the various operating points possible but does not verify whether these operating points are stable or unstable.

Following the Kurokawa analysis, the stability of an operating point can be studied with the help of a perturbation analysis around the operating point. The perturbation is to be conducted up to the first order in terms of the currents i_{IN} and i_L; that is, equivalently up to the first order relative to the device admittance Y_{IN} and Y_L, respectively.

Let us again consider the admittance model of an oscillator as shown in Figure 7.3. The basic oscillator is still divided into a linear frequency-sensitive admittance $Y_L(\omega)$ and a nonlinear or device admittance $Y_{IN}(A, \omega)$, which is both frequency- and amplitude-sensitive. A perturbative (noise) current $i_N(t)$ with frequency around the fundamental frequency is now placed in parallel with the circuit to disrupt the ideal

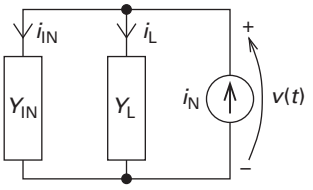


Figure 7.3 An admittance model of an oscillator.

steady-state operation. Perturbation from higher harmonics can also be considered (see Section 7.4). However, in conventional oscillators, the linear part usually represents a tank or resonator. The tank circuit behaves as an open at resonance and as a short for the harmonics, such that the dominant perturbations are usually around the fundamental frequency.[2]

The perturbed voltage $v(t)$ of the oscillator is then

$$v(t) = \mathrm{Re}\left\{A(t)e^{j[\omega_0 t+\phi(t)]}\right\} + \text{harmonics}, \tag{7.1}$$

where $A(t)$ and $\phi(t)$ are the instantaneous amplitude and phase, respectively, of the oscillation, which are now assumed to be slowly varying in time. This is the so-called envelope approximation. Note that ω_0 will be taken as the operating frequency in steady state.

To perform the stability and noise analysis, the Kurokawa theory introduces the concept of the instantaneous frequency $\omega_i(t)$ defined from the eigenvalue of the time derivative operator:

$$\frac{d}{dt}\left[A(t)e^{j[\omega_0 t+\phi(t)]}\right] = j\omega_i(t)\left[A(t)e^{j(\omega_0 t+\phi(t))}\right],$$

where the instantaneous frequency ω_i is given by

$$\omega_i(t) = \omega_0 + \frac{d\phi(t)}{dt} - j\frac{1}{A(t)}\frac{dA(t)}{dt} = \omega_0 + \delta\omega(t).$$

The current $i_{IN}(t)$ in the nonlinear device can still be expressed as

$$i_{IN}(t) = \mathrm{Re}\left\{A(t)e^{j[\omega_0 t+\phi(t)]}Y_{IN}[A(t), \omega_i(t)]\right\} + \text{harmonics}. \tag{7.2}$$

Note that, instead of ω_0 and A_0, the instantaneous amplitude $A(t)$ and instantaneous frequency $\omega_i(t)$ are now used in the functional dependence of Y_{IN}. Similarly, the current i_L flowing through the linear part is still given by

$$i_L(t) = \mathrm{Re}\left\{A(t)e^{j[\omega_0 t+\phi(t)]}Y_L[\omega_i(t)]\right\} + \text{harmonics}, \tag{7.3}$$

where we use $\omega_i(t)$ instead of ω_0 in Y_L.

[2] The assumption that Y_L represents a tank is not necessary. The test circuit in Figure 7.2 can handle large voltage harmonics. Ring oscillators can also be analyzed with a modified test circuit (to be reported elsewhere).

Owing to the current conservation, the perturbation current $i_N = i_L + i_{IN}$ satisfies

$$i_N(t) = \mathrm{Re}\left\{A(t)e^{j[\omega_0 t+\phi(t)]}\left(Y_L[\omega_i(t)] + Y_{IN}[A(t), \omega_i(t)]\right)\right\} + \text{harmonics}. \tag{7.4}$$

We shall now assume that the quantities $d\phi/dt$ and $(1/A)dA/dt$ in ω_i are much smaller than ω_0. By performing a Taylor series expansion in both Y_L and Y_{IN} in terms of $\delta\omega(t)$, multiplying then both sides of Equation (7.4) by $\cos(\omega_0 t + \phi(t))$ and $-\sin(\omega_0 t + \phi(t))$ and finally integrating each of those equations in time over one period of oscillation T, we obtain the following relations:

$$\begin{aligned} i_{N1}(t) &= \frac{2}{T}\int_{t-T}^{t} i_N \cos[\omega_0 t + \phi(t)]dt \\ &= A(t)\left(G_L(\omega_0) + G_{IN}[A(t), \omega_0] + G'_T(A_0, \omega_0)\frac{d\phi}{dt} + \frac{B'_T(A_0, \omega_0)}{A(t)}\frac{dA}{dt}\right), \end{aligned} \tag{7.5}$$

$$\begin{aligned} -i_{N2}(t) &= -\frac{2}{T}\int_{t-T}^{t} i_N \sin[\omega_0 t + \phi(t)]dt \\ &= A(t)\left(B_L(\omega_0) + B_{IN}[A(t), \omega_0] + B'_T(A_0, \omega_0)\frac{d\phi}{dt} - \frac{G'_T(A_0, \omega_0)}{A(t)}\frac{dA}{dt}\right), \end{aligned} \tag{7.6}$$

where we define

$$\begin{aligned} G'_T(A_0, \omega_0) &= G'_L(\omega_0) + G'_{IN}(A_0, \omega_0) = \left.\frac{\partial G_L}{\partial \omega}\right|_{\omega_0} + \left.\frac{\partial G_{IN}}{\partial \omega}\right|_{A_0,\omega_0}, \\ B'_T(A_0, \omega_0) &= B'_L(\omega_0) + B'_{IN}(A_0, \omega_0) = \left.\frac{\partial B_L}{\partial \omega}\right|_{\omega_0} + \left.\frac{\partial B_{IN}}{\partial \omega}\right|_{A_0,\omega_0}. \end{aligned}$$

Note that under steady-state conditions ($dA/dt = d\phi/dt = 0$, $\omega_i = \omega_0$, $A = A_0$, and $i_N = 0$) we have

$$\begin{aligned} G_L(\omega_0) + G_{IN}(A_0, \omega_0) &= 0, \\ B_L(\omega_0) + B_{IN}(A_0, \omega_0) &= 0. \end{aligned}$$

Equations (7.5) and (7.6) can be further simplified by noting that, for small δA relative to A_0, we have the following perturbative expansion:

$$\begin{aligned} G_{IN}[A(t), \omega_0] &= G_{IN}(A_0, \omega_0) + \frac{\partial G_{IN}(A_0, \omega_0)}{\partial A}\delta A \\ &= G_{IN}(A_0, \omega_0) + G'_{IN}(A_0, \omega_0)\delta A, \\ B_{IN}[A(t), \omega_0] &= B_{IN}(A_0, \omega_0) + \frac{\partial B_{IN}(A_0, \omega_0)}{\partial A}\delta A \\ &= B_{IN}(A_0, \omega_0) + B'_{IN}(A_0, \omega_0)\delta A. \end{aligned}$$

Hence Equations (7.5) and (7.6) can be simplified as

$$\frac{i_{\mathrm{N1}}(t)}{A_0} = \frac{\partial G_{\mathrm{IN}}(A_0,\omega_0)}{\partial A}\,\delta A + G'_{\mathrm{T}}(A_0,\omega_0)\frac{d\phi}{dt} + \frac{B'_{\mathrm{T}}(A_0,\omega_0)}{A_0}\frac{d\delta A}{dt},$$
$$-\frac{i_{\mathrm{N2}}(t)}{A_0} = \frac{\partial B_{\mathrm{IN}}(A_0,\omega_0)}{\partial A}\,\delta A + B'_{\mathrm{T}}(A_0,\omega_0)\frac{d\phi}{dt} - \frac{G'_{\mathrm{T}}(A_0,\omega_0)}{A_0}\frac{d\delta A}{dt}.$$

The above two equations can be rewritten into the following final master equations:

$$\left|Y'_{\mathrm{T}}(\omega_0)\right|^2\,\frac{d\delta A(t)}{dt} + A_0\beta\;\delta A(t) = B'_{\mathrm{T}}(A_0,\omega_0)i_{\mathrm{N1}}(t) + G'_{\mathrm{T}}(A_0,\omega_0)i_{\mathrm{N2}}(t), \tag{7.7}$$

$$A_0\left(\left|Y'_{\mathrm{T}}(\omega_0)\right|^2\,\frac{d\phi(t)}{dt} + \alpha\;\delta A(t)\right) = G'_{\mathrm{T}}(A_0,\omega_0)i_{\mathrm{N1}}(t) - B'_{\mathrm{T}}(A_0,\omega_0)i_{\mathrm{N2}}(t), \tag{7.8}$$

where we define Y'_{T} as

$$Y'_{\mathrm{T}} = G'_{\mathrm{T},0} + jB'_{\mathrm{T},0} = G'_{\mathrm{T}}(A_0,\omega_0) + jB'_{\mathrm{T}}(A_0,\omega_0) = \left.\frac{\partial Y_{\mathrm{L}}}{\partial\omega}\right|_{\omega_0} + j\left.\frac{\partial Y_{\mathrm{IN}}}{\partial\omega}\right|_{A_0,\omega_0},$$

and the correlation factor α and stability factor β for shunt resonance as

$$\alpha = G'_{\mathrm{T}}(A_0,\omega_0)G'_{\mathrm{IN}}(A_0,\omega_0) + B'_{\mathrm{T}}(A_0,\omega_0)B'_{\mathrm{IN}}(A_0,\omega_0),$$
$$\beta = B'_{\mathrm{T}}(A_0,\omega_0)G'_{\mathrm{IN}}(A_0,\omega_0) - G'_{\mathrm{T}}(A_0,\omega_0)B'_{\mathrm{IN}}(A_0,\omega_0).$$

These equations are the same as those originally derived by Kurokawa [1] except for the inclusion of the frequency dependence of $Y_{\mathrm{IN}}(A,\omega)$ in the coefficient Y'_{T} replacing Y'_{L}.

It results from the master equation (7.7) that, for the operation point (A_0,ω_0) to be stable, the perturbation $\delta A(t)$ must decay, and therefore the stability factor must be positive:

$$\beta > 0.$$

Note that the factor β can be expressed in terms of the following generalized scalar cross product:

$$\beta = \mathbf{Y}'_{\mathrm{IN},0} \times \mathbf{Y}'_{\mathrm{T},0} = \left|Y'_{\mathrm{T},0}\right|\left|Y'_{\mathrm{IN},0}\right|\sin\theta,$$

where we define the vectors $\mathbf{Y}'_{\mathrm{T},0} = \left[G'_{\mathrm{T},0}\;B'_{\mathrm{T},0}\right]$ and $\mathbf{Y}'_{\mathrm{IN},0} = \left[G'_{\mathrm{IN},0}\;B'_{\mathrm{IN},0}\right] = \left[G'_{\mathrm{IN}}(A_0,\omega_0)\;B'_{\mathrm{IN}}(A_0,\omega_0)\right]$. The angle θ is defined, for consistency with the Kurokawa theory [1] [2], as the angle clockwise from the device line direction $-\mathbf{Y}'_{\mathrm{IN},0}$ to the load line direction $\mathbf{Y}'_{\mathrm{T}}$ as shown in Figure 7.4. The stability coefficient β is then positive if we have

$$\sin\theta > 0.$$

It results that the angle θ must be between 0° and 180° in order for the operating point to be stable. Graphical examples for a stable and an unstable operating point are shown in Figures 7.4(a) and (b), respectively.

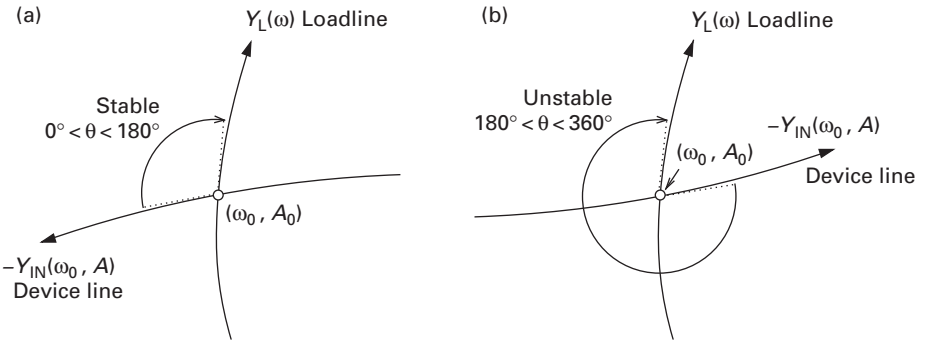


Figure 7.4 Stable (a) and unstable (b) operating points using Kurokawa graphical analysis.

7.3 Vector measurement of device line with real-time active load-pull

Oscillators are usually designed using circuit simulations. In harmonic-balance simulators, joint power sweep (A) and frequency sweep (ω) are used to determine the operating point (A_0, ω_0) which enforces the unity loop gain $\Gamma_L(\omega)\Gamma_{IN}(A, \omega) = 1$. However, the reliability of the simulation results obtained for predicting the performance of a physical circuit realization will depend on the fidelity of the device and circuit models. It results that in many situations it would be desirable to be able to measure the device line Y_{IN} to assist with the design of the oscillator circuit. The device line can also be used in turn to optimize the output power and the phase-noise characteristics of the oscillator [1] [3]. By finding the optimal load network of the oscillator at the fundamental frequency, the output power can be maximized [4] [5] [6]. In addition, the impedance termination for the harmonics can be used to optimize the output power, phase noise, and output harmonic content of the oscillator. Results from a previous study using active load-pull measurements [7] indicate that the second-harmonic load impedance is of importance for optimizing the oscillator's output power and efficiency. In this section we shall discuss the use of power sweep and real-time active load-pull (RTALP) for the rapid prototyping of a 2.5-GHz discrete oscillator.

7.3.1 Test oscillator circuit

The oscillator circuit used for our analysis is shown in Figure 7.5. It is realized with an ATF54143 pHEMT. A series feedback network and a terminating network are used to induce a stable input reflection coefficient $\Gamma_{IN}(|a_1|, \omega_0)$ with magnitude larger than unity at the targeted frequency of 2.5 GHz [8]. For the associated negative resistance not to break into self-oscillation during the measurements, the Nyquist stability condition needs to be satisfied. That is, care must be taken that the loadline does not circle the device line. As we shall see, this is readily achieved for a source impedance of 50 Ω if the device line stays clear of the region neighboring the center of the Smith chart.

A drain voltage of 2.0 V and a gate voltage of 0.55 V yielding a drain current of 27 mA are applied to the device for its DC biasing. Both a broadband bias tee and a $\lambda/4$ bias line are used for the drain biasing.

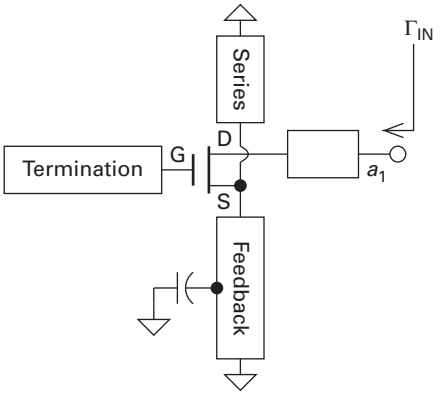


Figure 7.5 A schematic diagram of the resonant negative-resistance circuit tested.

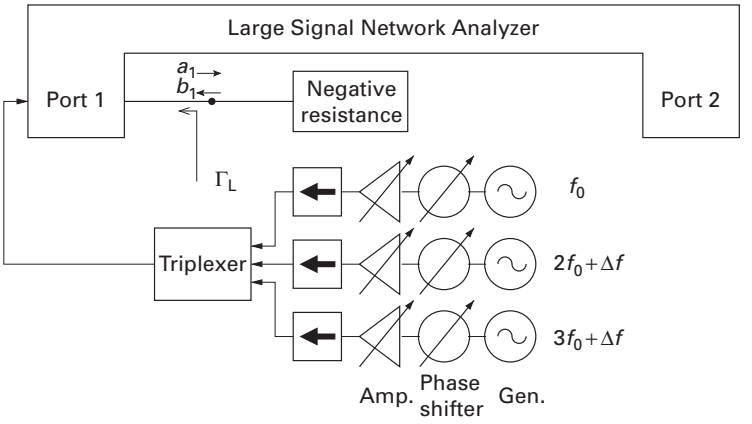


Figure 7.6 RTALP system used for the multi-harmonic characterization of the nonlinear negative resistance of an oscillator designed to operate at the frequency f.

The terminating network is implemented with a transmission-line resonator. Series feedback is implemented using a shorted stub with a capacitor tap for tuning the negative resistance.

7.3.2 Real-time multi-harmonic active load-pull system

The diagram of the testbed used for the RF measurements is shown in Figure 7.6. This test system which is configured with an LSNA is used to perform the frequency sweep, power sweep, active load-pull, and real-time multi-harmonic active load-pull.

As shown, the RF sources $(n\omega_0)$ are connected to the oscillator via port 1 of the LSNA. The incident wave $a_1(n\omega_0)$ injected from the RF source $(n\omega_0)$ and the reflected wave $b_1(n\omega_0)$ from the oscillator are then measured with the LSNA.

To determine the optimal impedance termination for the second and third harmonics, we adopt the the multi-harmonic RTALP technique introduced in Chapter 5. For

this measurement, a frequency offset $\Delta\omega$ of about 200 kHz is used. In this method an incident wave (see Figure 7.6) at $n\omega_0 + \Delta\omega$ is injected into the oscillator network. The time-varying reflection coefficient obtained for a given incident power,

$$P_{\text{inc}} = (n\omega_0 + \Delta\omega) = \frac{1}{2}|a_1(n\omega_0 + \Delta\omega)|^2, \tag{7.9}$$

can be calculated using

$$\Gamma_{\text{L}}(n\omega_0, t) = \frac{\sum_{p=-\text{SSB}}^{\text{SSB}} a_1(n\omega_0 + p\,\Delta\omega)e^{jp\Delta\omega t}}{\sum_{p=-\text{SSB}}^{\text{SSB}} b_1(n\omega_0 + p\,\Delta\omega)e^{jp\Delta\omega t}}.$$

In addition to the reflection coefficient, the total output power can be represented by

$$P_{\text{out,total}} = \sum_{n=1}^{N} P_{\text{out}}(n\omega_0, t),$$

where we defined $P_{\text{out}}(n\omega_0, t)$ as

$$P_{\text{out}}(n\omega_0, t) = -\frac{1}{2}\sum_{p=-\text{SSB}}^{\text{SSB}}\sum_{q=-\text{SSB}}^{\text{SSB}} v_1(n\omega_0 + p\,\Delta\omega)i_1^*(n\omega_0 + q\,\Delta\omega)e^{j(p-q)\Delta\omega t}. \tag{7.10}$$

7.3.3 Experimental results

The amplitude of the small-signal input reflection coefficient $|\Gamma_{\text{IN}}(0, \omega_0)|$ versus frequency measured from 2 GHz to 3 GHz with the LSNA in its network-analyzer mode is shown in Figure 7.7. The tuning capacitor and the length of the stubs were adjusted such that the negative resistance peaked within the desired frequency range. As a result, a magnitude of 3.7 is observed at 2.5 GHz in Figure 7.7.

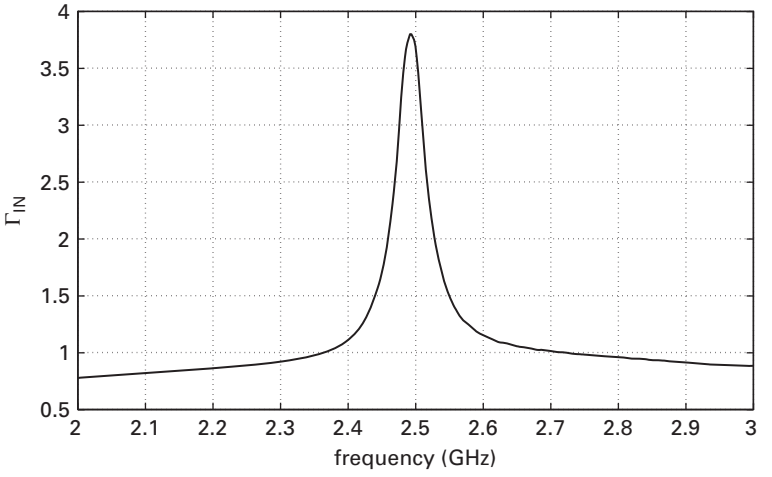


Figure 7.7 Measured magnitude of $\Gamma_{\text{IN}}(0, \omega)$ versus frequency. (From [8] with permission, ©2007 IEEE.)

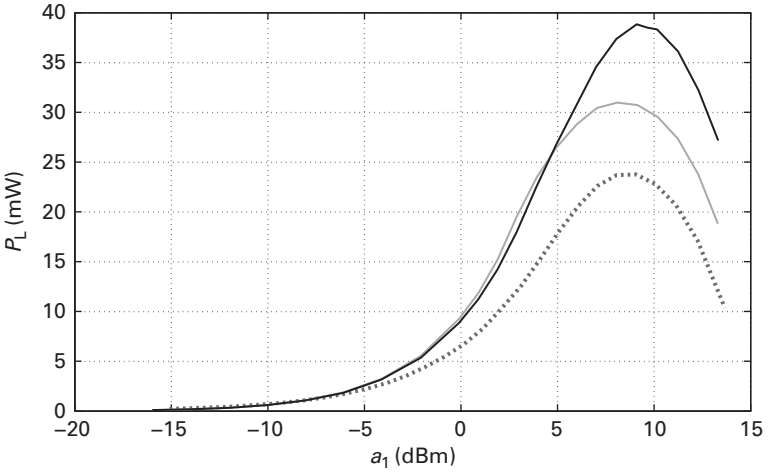


Figure 7.8 Comparison of output power versus incident power for various biasing configurations and second-harmonic impedance termination. A maximum output power ($P_{L,max}$) of 23.8 mW (dotted line) is obtained with the $\lambda/4$ drain bias tee. With a broadband drain bias tee, $P_{L,max}$ increases to 31.0 mW (gray line). Using both recursive RTALP and power sweep, a maximum output power of 38.8 mW is obtained with 9.1 dBm a_1 (black line). (From [8] with permission, ©2007 IEEE.)

Large-signal measurements are performed next. The nonlinear input reflection coefficient can be obtained from both the incident wave a_1 and the reflected wave b_1 measured by the LSNA, using

$$\Gamma_{IN}(|a_1|, \omega) = \frac{b_1(\omega)}{a_1(\omega)}. \tag{7.11}$$

The output power delivered to the load can be then calculated using

$$P_L(\omega_0) = \frac{1}{2}|b_1(\omega_0)|^2 - \frac{1}{2}|a_1(\omega_0)|^2 \tag{7.12}$$

To increase the maximum output power from the oscillator, the incident power on the negative resistance is swept. A comparison of the output power obtained versus incident power for two different biasing configurations is shown in Figure 7.8. The output power at the fundamental frequency obtained by using a $\lambda/4$ high-impedance bias line connected to the drain is shown (dotted line). A maximum output power of 23.8 mW is obtained. For this measurement, a 50-Ω termination is used for the second-harmonic load impedance. Since the $\lambda/4$ bias line provides an open termination at the fundamental frequency, it also implements a short at the second harmonic, preventing any further second-harmonic optimization from the load circuit. To circumvent this problem, a broadband bias tee can be used so that the second harmonic can be productively tuned. As can be seen in Figure 7.8, the impact of the bias tee is quite significant as the output power is increased to 31.0 mW. The load impedance at the fundamental frequency which provides the maximum output power $P_{L,max}$ can then be identified using

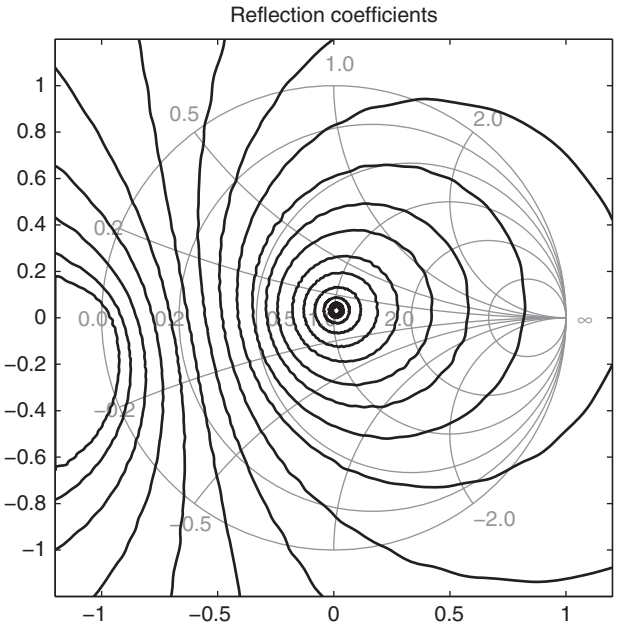


Figure 7.9 Loci of $\Gamma_L(2\omega_0, t)$ obtained from the RTALP measurement with the LSNA. A frequency offset of about 200 kHz is used. (From [8] with permission, ©2007 IEEE.)

$$\Gamma_L(\omega_0) = \frac{1}{\Gamma_{IN}(|a_{1,opt}|, \omega_0)}. \tag{7.13}$$

While keeping this optimum $\Gamma_L(\omega_0)$, the RTALP technique is applied next to find the optimum $\Gamma_L(2\omega_0)$ which provides maximum output power. For this measurement, a $2\omega_0 + \Delta\omega$ signal is injected from the RF source ($2\omega_0$) and a 2.5-GHz signal is injected from the RF source (ω_0). Figure 7.9 shows the loci of $\Gamma_L(2\omega_0, t)$ obtained from the RTALP measurements.

The output power contour plot in the $\Gamma_L(2\omega_0)$ plane is depicted in Figure 7.10. This output power contour plot is generated on the basis of Equation (7.10) with the loci of $\Gamma_L(2\omega_0)$. As can be seen, the maximum output power of 39.5 mW is obtained. However, this is slightly affected by memory effects due to the phase sweeping. For the same conditions an output power of 38.3 mW has been verified to be actually obtained using active load-pull.

By fixing the second-harmonic load impedance to its optimal value, we can remeasure a new output power versus incident power, which now accounts for the second-harmonic termination. As shown in Figure 7.8, the maximum output power of the oscillator is increased in the end from 31.0 mW to 38.8 mW. A new optimum value of $\Gamma_L(\omega_0)$ that provides this maximum output power is also obtained from this measurement. A comparison of the output powers obtained for active load-pull and constant phase under various measurement conditions is given in Table 7.1.

The optimum third-harmonic load impedance was also determined using RTALP measurement. However, the third harmonic was found to have a negligible impact on the

Table 7.1. Comparison of the measured output powers. (From [8] with permission, ©2007 IEEE.)

Method ω_0	Termination $\Gamma_L(\omega_0)$	Method $2\omega_0$	Termination $\Gamma_L(2\omega_0)$	Output power (mW)
Power sweep	$0.415\angle 177.8°$	Load	$0\angle 0.0°$	31.0
CW power	$0.415\angle 177.8°$	RTALP	$1\angle 168.0°$	39.5
CW power	$0.415\angle 177.8°$	ALP	$1\angle 168.0°$	38.3
Power sweep	$0.416\angle 171.0°$	ALP	$1\angle 168.0°$	38.8

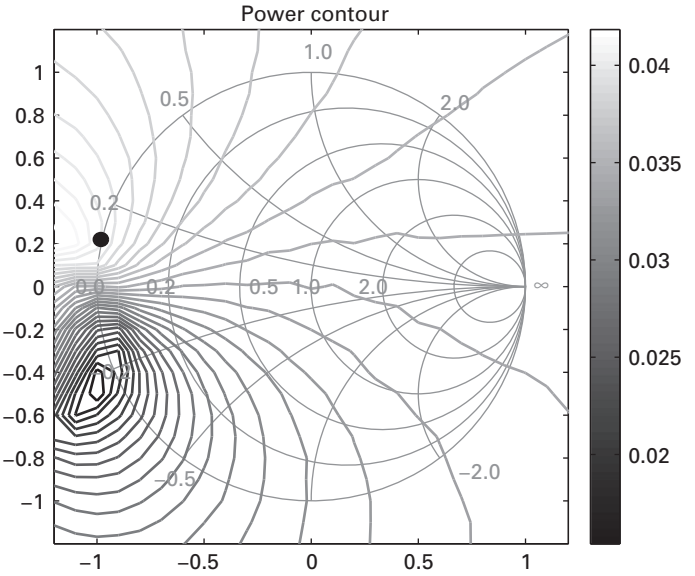


Figure 7.10 Output power contour plot in the $\Gamma_L(2\omega_0)$ plane, obtained from the RTALP measurement by the LSNA. The black dot indicates the optimum second-harmonic load impedance ($\Gamma_{L,opt}(2\omega_0) = 1\angle 168°$) which maximizes the output power (39.5 mW). (From [8] with permission, ©2007 IEEE.)

maximum oscillator output power. Other oscillator results with third-harmonic output results are presented in Ref. [9].

A trajectory plot of the device line $\Gamma_{IN}^{-1}(|a_1(\omega_0)|, \omega_0)$ for increasing incident power $|a_1(\omega_0)|$, obtained with (black top line) and without (gray bottom line) the optimum $\Gamma_{L,opt}(2\omega_0)$ termination, is shown in Figure 7.11. The black arrow indicates the direction of increased incident power levels. The power is swept from -15 dBm to 14 dBm in steps of 1 dBm. The bottom gray line is initially obtained from the power sweep using a load termination of 50 Ω for the second harmonic. The bottom gray dot gives the optimum load $\Gamma_{L,opt1}(\omega_0) = 0.415 \angle 177.84°$ which provides the maximum output power of 31.0 mW for 8.1 dBm of incident power. The top black line is measured after applying the optimum second-harmonic termination $\Gamma_{L,opt}(2\omega_0)$ obtained from the RTALP measurements. The black dot locates the optimum load $\Gamma_{L,opt2}(\omega_0) = 0.416 \angle 171.0°$ which

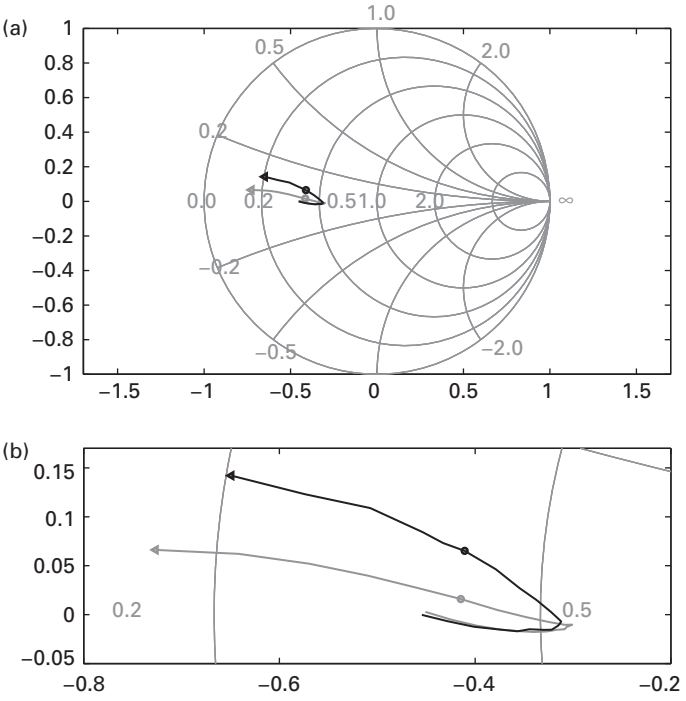


Figure 7.11 Device line $\Gamma_{IN}^{-1}(|a_1|, \omega_0)$ versus the incident power $P_{inc} = (1/2)|a_1(\omega_0)|^2$ for two different measurement conditions. (From [8] with permission, ©2007 IEEE.)

provides the maximum output power of 38.8 mW with the incident power of 9.1 dBm. As can be seen, the trajectory of the device loadline is slightly modified when using the optimum load termination $\Gamma_{L,opt}(2\omega_0)$ for the second harmonic.

7.3.4 Self-oscillation test

To verify the accuracy of the device line $\Gamma_{IN}^{-1}(|a_1|, \omega_0)$ for realizing an oscillator using the Kurokawa theory, a self-oscillating circuit was tested [8]. A schematic diagram of the experimental testbed used for the self-oscillation measurement is shown in Figure 7.12. The broadband bias tee of the LSNA is used for the drain bias in this test. The total loss including tuner (0.8 dB) and bias tee plus LSNA coupler plus cable (6.9 dB) losses adds up to 7.7 dB.

A load tuner is used to set the targeted reflection coefficient at 2.5 GHz. However, the impedance of the second harmonic could not be controlled in this self-oscillating testbed but was verified to be around $\Gamma_L(2\omega_0) = 0.43 \angle 100°$. The device line $\Gamma_{IN}^{-1}(|a_1|, \omega_0)$ is therefore measured for this constant second-harmonic termination.

Both the loadline $\Gamma_L(\omega_0)$ and the device line $\Gamma_{IN}^{-1}(|a_1|, \omega_0)$ are plotted in Figure 7.13. In this measurement, $\Gamma_L(\omega_0)$ is swept from 2 GHz to 3 GHz (gray arrow). The device line is measured with the LSNA at 2.5 GHz with incident power from −15 dBm to 14 dBm (black line) while keeping $\Gamma_L(2\omega_0) = 0.43 \angle 100°$, which is the approximately $2\omega_0$

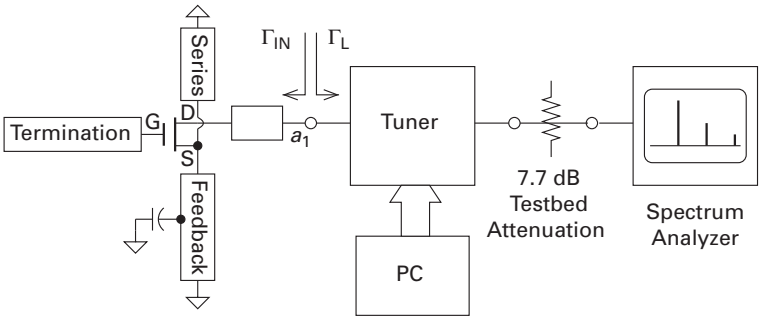


Figure 7.12 A schematic diagram of the testbed used for verifying the self-oscillation. (From [8] with permission, ©2007 IEEE.)

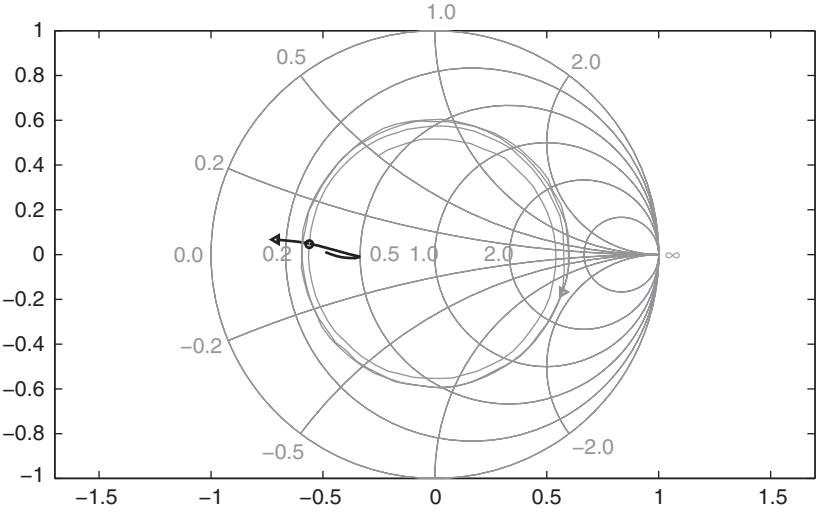


Figure 7.13 Loci of the device line $\Gamma_{IN}^{-1}(|a_1|, \omega_0)$ and the loadline $\Gamma_L(\omega_0)$ of the tuner used to realize the self-oscillating oscillator. (From [8] with permission, ©2007 IEEE.)

termination provided by the tuner. The black dot indicates the expected operating point at $\Gamma_L(\omega_0) = \Gamma_{IN}^{-1}(|a_1|, \omega_0) = 0.56 \angle 175°$ of the oscillator at 2.5 GHz. This operating point yields a power of 28.3 mW (14.5 dBm) as shown in Figure 7.14. The operating point satisfied the graphical stability criteria ($0° < \theta < 180°$). Note that the Smith chart provides a conformal mapping of the admittance (Y) plane in the reflection-coefficient (Γ) plane, such that the angle between any two direction vectors is conserved in the Y-to-Γ mapping. The same Kurokawa rules as those derived in the Y plane can therefore also be applied for the analysis of the stability of the operating point in the Γ plane.

Finally, note that the Nyquist requirement at $|a_1| = 0$ for starting the oscillation with the load tuner forces the selection of an operating point below the maximum-power point in Figure 7.14.

Self-oscillation was verified using the load tuner and a spectrum analyzer. A comparison of predicted results from Kurokawa theory using LSNA measurements and

Table 7.2. Comparison of predicted and measured oscillation frequency and harmonic power. (From [8] with permission, ©2007 IEEE.)

	Kurokawa theory	Measured
Fundamental frequency (GHz)	2.5	2.515
Fundamental power (dBm)	14.5	14.1
Second-harmonic power (dBm)	1.22	−3.19
Third-harmonic power (dBm)	−29.08	−36.8

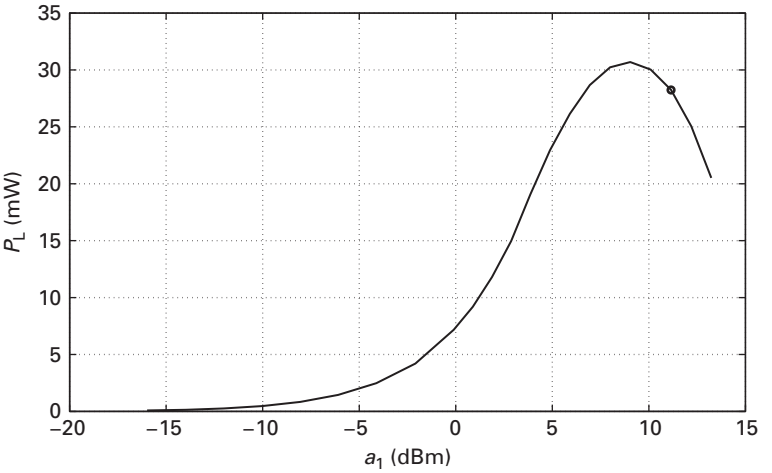


Figure 7.14 Output power versus a_1 calculated from LSNA measurements. The dot gives the predicted output power of 28.3 mW (14.5 dBm) at the operating point $|a_1|$ of 11.2 dBm obtained in Figure 7.13 at 2.5 GHz. (From [8] with permission, ©2007 IEEE.)

spectrum-analyzer measurements of the self-oscillating oscillator circuit is summarized in Table 7.2. The observed frequency of self-oscillation is 2.515 GHz. Accounting for the losses from the testbed (6.9 dB) and tuner (0.8 dB), an oscillator output power of 14.075 dBm is obtained. This output power is in reasonable agreement with the predicted output power of 14.5 dBm in Figure 7.14, considering that the second-harmonic load impedance is controlled only approximately in this simple self-oscillating testbed. A complete circuit realization of the oscillator circuit with a multi-harmonic load circuit for optimal output power is reported in Ref. [9]. Note that, as we shall see in the next sections, an optimal output power is an important criterion for reducing the output phase noise of an oscillator in the case of white noise. In the case of $1/f$ noise only the amplitude noise will be reduced with increasing output power.

7.4 Impact of white noise on an oscillator

In this section we shall discuss the application of the Kurokawa theory to the derivation of amplitude and phase noise in an oscillator in the presence of white noise. The

Kurokawa theory provides valuable circuit insights that complement the computational noise-analysis techniques implemented in modern harmonic-balance simulators.

The advantage of the circuit-based Kurokawa theory over more recently developed phase-noise theories such as the impulse sensitivity function (ISF) approach [14] is its ability to account for the correlation between amplitude and phase noise (AM-to-PM conversion). Yet in the uncorrelated case (AM-to-PM conversion neglected) the generalized Kurokawa theory yields similar results to the ISF approach and harmonic-balance simulations. For the correlated case (strong AM-to-PM conversion) the generalized Kurokawa theory yields similar results to harmonic-balance-based simulations that use the conversion-matrix method [15] [16] and the perturbation projection vector (PPV) technique [10] [11] [12] [13].

However, unlike other approaches [10] [11] [12] that rely on numerical computations, the Kurokawa theory yields analytic circuit-based results that can be easily used by circuit designers. In the generalized Kurokawa model presented here, the noise sources can be either stationary (white noise) or cyclostationary (up-converted $1/f$ noise; down-converted white noise or shot noise associated with an RF current). Note that a cyclostationary noise can be decomposed as the product of a stationary noise source and a deterministic periodic RF oscillation. Additional cyclostationary effects such as the impact of large-signal RF oscillations on noise sources can also be accounted for, as was discussed in Chapter 5, for the Lorentzian (popcorn) and $1/f$ noises generated by traps [17].

The master equations (7.7) and (7.8) derived in Section 7.1 are still applicable except that now the perturbing current i_{N} applied across the oscillator is a white-noise source. Flicker noise will be discussed in the next section.

Following Kurokawa [1], we have the following spectral densities:[3]

$$\overline{S_{i_{\mathrm{N}}}(f)} = |e|^2, \qquad S_{i_{\mathrm{N1}}}(f) = S_{i_{\mathrm{N2}}}(f) = 2|e|^2.$$

If we assume that Equations (7.7) and (7.8) hold for all time (oscillator on for a long time) and $i_{\mathrm{N1}}(t)$ and $i_{\mathrm{N2}}(t)$ are stationary processes then $A(t)$ and $\phi(t)$ are also themselves stationary [18]. On taking the Fourier transform of Equations (7.7) and (7.8), we get

$$j\,\Delta\omega \frac{\delta A(\Delta\omega)}{A_0}\left|Y'_{\mathrm{T}}\right|^2 + \beta\,\delta A(\Delta\omega) = \frac{1}{A_0}\left[i_{\mathrm{N1}}(\Delta\omega)B'_{\mathrm{T},0} + i_{\mathrm{N2}}(\Delta\omega)G'_{\mathrm{T},0}\right], \quad (7.14)$$

$$j\,\Delta\omega\left|Y'_{\mathrm{T}}\right|^2\phi(\Delta\omega) + \alpha\,\delta A(\Delta\omega) = \frac{1}{A_0}\left[i_{\mathrm{N1}}(\Delta\omega)G'_{\mathrm{T},0} - i_{\mathrm{N2}}(\Delta\omega)B'_{\mathrm{T},0}\right]. \quad (7.15)$$

Note that in this Fourier transform we use the frequency $\Delta\omega$ to emphasize that we are dealing with a small (usually below 10 MHz) offset frequency compared with the RF frequency ω_0. From Equation (7.14) we get the amplitude-variation spectral density

[3] This is equivalent to the so-called narrowband representation of noise: $i_{\mathrm{N}} = i_{N1}\cos(\omega_0 t + \varphi) + i_{\mathrm{N2}}\sin(\omega_0 t + \varphi)$. Since the RF noise i_{N} has twice the bandwidth B of the baseband noises i_{N1} and i_{N2}, we have $\langle i_{\mathrm{N}}^2\rangle = 2BS_{i_{\mathrm{N}}} = BS_{i_{\mathrm{N1}}} = BS_{i_{\mathrm{N2}}} = \langle i_{\mathrm{N1}}^2\rangle = \langle i_{\mathrm{N2}}^2\rangle$.

$$S_{\delta A}(\Delta\omega) = \frac{2\left|Y'_{\mathrm{T}}(\omega_0)\right|^2 |e|^2}{A_0^2\beta^2 + \Delta\omega^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^4},$$

using $|i_{\mathrm{N1}}(\Delta\omega)|^2 = |i_{\mathrm{N2}}(\Delta\omega)|^2 = 2|e|^2$ [1]. Then $S_{\delta A}(\Delta\omega)$ is normalized by A_0^2 to obtain the following final analytic expression for the white amplitude noise:

$$S_{a,\mathrm{white}}(\Delta\omega) = \frac{2\left|Y'_{\mathrm{T}}(\omega_0)\right|^2 |e|^2}{A_0^4\beta^2 + \Delta\omega^2 A_0^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^4}. \tag{7.16}$$

The autocorrelation $R_{\delta A}(\tau)$ is obtained from a Fourier transform of $S_{\delta A}(\Delta\omega)$:

$$R_{\delta A}(\tau) = \frac{|e|^2}{A_0\beta}\exp(-\eta|\tau|), \tag{7.17}$$

where we introduced the corner frequency η:

$$\eta = \frac{A_0\beta}{\left|Y'_{\mathrm{T}}(\omega_0)\right|^2}.$$

From Equation (7.15), we get the following expression for $S_{\phi,\mathrm{white}}$:

$$S_{\phi,\mathrm{white}}(\Delta\omega) = \frac{2|e|^2}{A_0^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^2\Delta\omega^2}\left(1+\frac{\alpha^2}{\beta^2}\right) - \frac{\left|Y'_{\mathrm{T}}(\omega_0)\right|^2}{\Delta\omega^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^4 + A_0^2\beta^2} \times \frac{2|e|^2\alpha^2}{A_0^2\beta^2}. \tag{7.18}$$

The phase double-sided spectral density obtained in Equation (7.18) is in agreement with [2] for the uncorrelated δA and ϕ case ($\alpha/\beta = 0$). Note that the general solution of Equation (7.15) is valid up to an arbitrary impulse:

$$S_{\phi,\mathrm{WHITE}}(\Delta\omega) = S_{\phi,\mathrm{white}}(\Delta\omega) + 2\pi\left(\sigma^2 + \frac{\alpha^2|e|^2}{\beta^3 A_0^3}\right)\delta(\omega). \tag{7.19}$$

The impulse weight is selected so as to satisfy the non-zero boundary condition of $R_\phi(0) = \sigma^2$. By taking the inverse Fourier transform of Equation (7.19) we obtain for the correlated case

$$R_\phi(\tau) = \sigma^2 - \frac{|e|^2}{A_0^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^2}\left(1+\frac{\alpha^2}{\beta^2}\right)|\tau| - \frac{\alpha^2|e|^2}{\beta^3 A_0^3}\big[\exp(-\eta|\tau|) - 1\big]. \tag{7.20}$$

Here we have used the relation $\alpha^2 + \beta^2 = \left|Y'_{\mathrm{IN}}(A_0,\omega_0)\right|^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^2$. These results have been computed using the following inverse Fourier transform pairs:

$$\mathcal{F}^{-1}\left[\frac{1}{\omega^2}\right] = -\frac{1}{2}|\tau| \qquad \text{and} \qquad \mathcal{F}^{-1}\left[\frac{1}{\xi^2+\Delta\omega^2}\right] = \frac{1}{2\xi}\exp(-\xi|\tau|).$$

In accordance with IEEE definitions [19], Equations (7.16) and (7.18) provide the amplitude noise $S_{a,\mathrm{white}}(\Delta\omega)$ and phase noise $\mathcal{L} = S_{\phi,\mathrm{white}}(\Delta\omega)$, for white noise injected across an oscillator with a shunt resonator. A phase detector is required to measure them. However, the amplitude and phase noises are sometimes evaluated from the

spectral analysis of the voltage noise using a spectrum analyzer, and it is useful to derive the voltage spectral density.

If the voltage across the tank is given by

$$v(t) = A(t)\cos[\omega_0 t + \phi(t)] + \text{harmonics},$$

then the autocorrelation of the voltage for stationary Gaussian noise processes is given by (see Appendix II in [3]),

$$R_V(\tau) = \frac{1}{2}\Big[A_0^2 + R_{\delta A}(\tau)\Big]\cos(\omega_0\tau)\exp\big[R_\phi(\tau) - R_\phi(0)\big]. \tag{7.21}$$

$R_{\delta A}(\tau)$ and $R_\phi(\tau)$ represent the autocorrelation functions of the amplitude deviation δA and the phase ϕ, respectively, derived in Equations (7.17) and (7.20).

The first term in Equation (7.21) which is proportional to A_0^2 is the PM noise. The second term involving $R_{\delta A}(\tau)$ is the AM noise term. Equation (7.21) is derived neglecting a third contribution arising from the correlation between phase and amplitude noises. This is justified since the PM noise will be verified below to be dominant over AM noise so that it can be assumed to be dominant also over this third contribution.

The autocorrelation function for the tank voltage is then

$$R_V(\tau) = \frac{1}{2}\left[A_0^2 + \frac{|e|^2}{A_0\beta}\exp(-\eta|\tau|)\right]\cos(\omega_0\tau) \times \exp\left[-\frac{|e|^2}{A_0^2\left|Y_T'(\omega_0)\right|^2}\left(1+\frac{\alpha^2}{\beta^2}\right)|\tau|\left(e^{-\eta|\tau|}-1\right)\right].$$

According to the Wiener–Khintchine theorem, the power spectral density of the voltage across the tank can now be obtained by taking the Fourier transform of $R_V(\tau)$:

$$S_V(\Delta\omega) = \mathcal{F}\{R_V(\tau)\}.$$

For $\alpha = 0$ (no correlation), a closed-form solution in the form of a Lorentzian (defined below) is obtained for S_V. No exact analytic solution is available for the Fourier transform of $R_V(\tau)$ when we account for the correlation between the amplitude and phase (α non-zero). However, an approximate analytic Lorentzian solution can be obtained for two practical limit conditions [3]:

$$S_{V,\text{ssb}}(\Delta\omega) = \underbrace{A_0^2\left[\frac{m_1}{m_1^2+\Delta\omega^2}\right]}_{\text{PM Noise}} + \underbrace{c\left[\frac{m_1+\eta}{(m_1+\eta)^2+\Delta\omega^2}\right]}_{\text{AM Noise}}, \tag{7.22}$$

where $S_{V,\text{ssb}}(\Delta\omega)$ is the single-sideband voltage spectral density. In Equation (7.22) we use $c = |e|^2/(A_0\beta)$ and define m_1 as

$$m_1 = \begin{cases} m_{01} & \text{for } \Delta\omega \gg \eta \text{ or for } \alpha = 0 \text{ and all } \Delta\omega, \\ m_{01}\left[1+\left(\dfrac{\alpha}{\beta}\right)^2\right] & \text{for } \eta \gg m_{01} \text{ and } \Delta\omega < \eta, \end{cases}$$

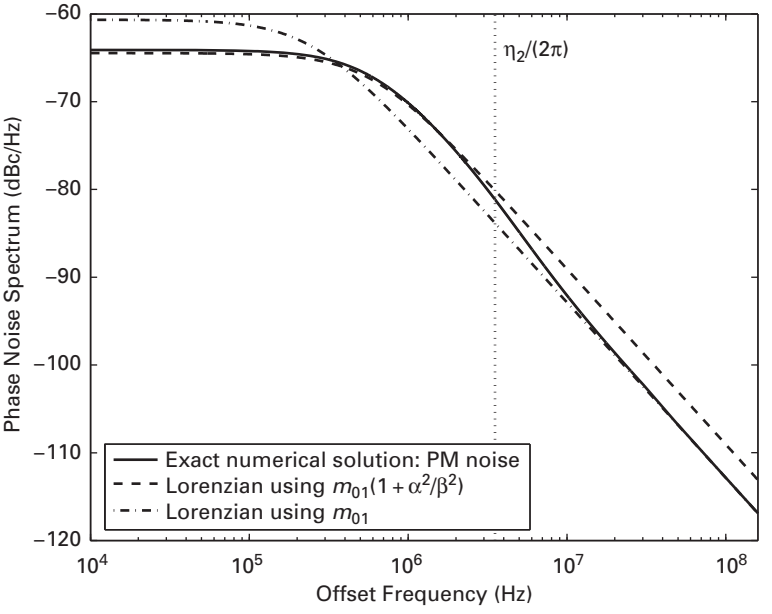


Figure 7.15 Comparison of PM voltage noise spectrum $S_{V,\mathrm{ssb}}(\Delta\omega)/\left(A_0^2/2\right)$ with various Lorentzian approximations for an *L–C* differential oscillator. (From [3] with permission, ©2007 IEEE.)

with

$$m_{01} = \frac{|e|^2}{A_0^2\left|Y_{\mathrm{T}}'(\omega_0)\right|^2}.$$

These asymptotic results therefore hold for large and small values of η. For intermediate value of η the Fourier transform of $R_V(\tau)$ can readily be calculated numerically. This is illustrated in Figure 7.15 for $\eta/(2\pi) = 3.5$ MHz. As is shown in Figure 7.15, $S_{V,\mathrm{ssb}}(\Delta\omega)/(A_0^2/2)$ is seen to relax for $\Delta\omega \gg \eta$ to the limiting Lorentzian with $\alpha = 0$. The voltage noise spectrum is no longer strictly a Lorentzian, and an inflexion point is introduced in the PM voltage noise density at the frequency $\eta/(2\pi)$.

Figure 7.16 compares the AM, PM, and AM-plus-PM white-noise components of $S_{V,\mathrm{ssb}}/(A_0^2/2)$ for a differential oscillator [3]. On a logarithmic graph the approximate PM/AM Lorentzian spectra have corner frequencies given by $m_1/(2\pi) \simeq 0.6$ MHz and $(m_1+\eta)/(2\pi) \simeq 4.2$ MHz, respectively. Since A_0^2 (PM noise) exceeds c (AM noise) (by 20 dB in Figure 7.16), $m_1/(2\pi)$ is the corner frequency of the total AM-plus-PM white-noise spectrum.

The inflexion point at η observed in the PM noise spectrum is not easily observable in the total AM-plus-PM noise spectrum due to the AM noise contribution. A close agreement of the theory with a circuit simulator (ADS) using the conversion-matrix method [15] [16] is verified at high offset frequencies ($\Delta\omega \gg m_1$) for the circuit considered. Note that the ADS simulator calculates only the IEEE phase noise, which grows indefinitely as the offset frequency vanishes.

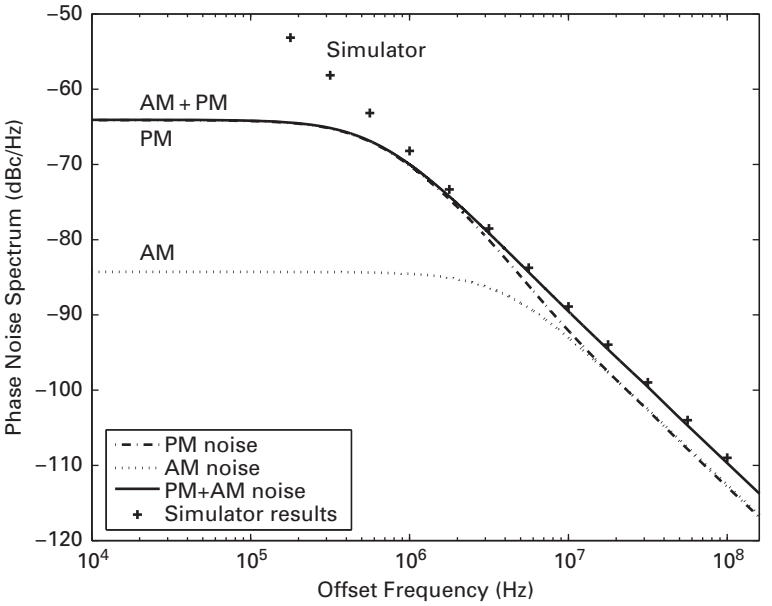


Figure 7.16 Comparison of the AM and PM voltage spectral density $S_{V,\mathrm{ssb}}(\Delta\omega)/\left(A_0^2/2\right)$ (plain and dotted lines) with ADS simulation using the IEEE phase noise $S_\phi(\Delta\omega)$ in an L–C differential oscillator. (From [3] with permission, ©2007 IEEE.)

Note that both m_1 and $m_1+\eta$ are proportional to $1/\left|Y'_{\mathrm{T}}(\omega_0)\right|^2$. For a parallel tank, $\left|Y'_{\mathrm{T}}(\omega_0)\right| \simeq 2C$ and is proportional to the tank Q. This shows that, at large offset frequencies, $S_V(\Delta\omega)$ is proportional to $1/Q^2$ [20]. Note that the equation derived above presents the voltage noise in terms of measurable parameters $Y'_{\mathrm{T}}(\omega_0)$ and $Y'_{\mathrm{IN}}(A_0,\omega_0)$.

The presence of the additional terms (α/β) in m_1 provides greater accuracy in the expression for the phase and voltage noise density. The ratio α/β should be as low as possible for reducing phase noise,

$$\frac{\alpha}{\beta}=\frac{\mathbf{Y}'_{\mathrm{IN}}\cdot\mathbf{Y}'_{\mathrm{T}}}{\mathbf{Y}'_{\mathrm{IN}}\times\mathbf{Y}'_{\mathrm{T}}}=\frac{\left|Y'_{\mathrm{T}}\right|\,\left|Y'_{\mathrm{IN}}\right|\cos\theta}{\left|Y'_{\mathrm{T}}\right|\,\left|Y'_{\mathrm{IN}}\right|\sin\theta}=\cot\theta,$$

where θ is the clockwise angle from $-\mathbf{Y}'_{\mathrm{IN}}$ to $\mathbf{Y}'_{\mathrm{T}}$. It results that when θ is 90° the noise correlation is minimized: $\alpha/\beta=0$. When θ is 0° or 180° the noise correlation is maximized: $\alpha/\beta=\infty$ [1]. In the circuit considered above, the correlation term is verified in Figure 7.15 to bring a shift on the order of ±3.8 dB for frequencies below $\eta/(2\pi)$.

The comparison of the Kurokawa theory presented above with the ISF theory [14] is also of interest. As suggested in Ref. [21], the ISF differential equation describing the phase evolution with time when a small perturbation signal $e(t)$ is applied can be defined as

$$\frac{d\phi(t)}{dt}=\Gamma[\omega_0 t+\phi(t)]e(t), \tag{7.23}$$

where $\Gamma(\theta)$ is the impulse sensitivity function which describes the response of the oscillation phase to a perturbation. The impulse sensitivity is periodic in θ with period 2π and can be expanded in a Fourier series:

$$\Gamma(\theta) = \mathrm{Re}\left\{\sum_{k=0}^{\infty} \Gamma_k \exp(jk\theta)\right\}.$$

The ISF equation (7.23) presents some similarities to the Kurokawa equation (7.8) except that the correlation of the phase to the amplitude fluctuation (AM to PM) is not accounted for.

Note that only the baseband frequency components of $\Gamma[\omega_0 t + \phi(t)]e(t)$ impact the slow phase evolution of $\phi(t)$, the higher RF oscillations of $\Gamma[\omega_0 t + \phi(t)]e(t)$ being averaged out. Therefore, in the ISF theory, the component of the white noise $e(t)$ at the fundamental frequency ω_0 is down converted to baseband frequencies by the Fourier coefficient Γ_1 of the fundamental frequency ($k = 1$ for oscillations at ω_0). Inspection of Equation (7.15) thus indicates that the coefficients $G'_{\mathrm{T},0}/\left|Y'_{\mathrm{T}}(\omega_0)\right|^2$ and $B'_{\mathrm{T},0}/\left|Y'_{\mathrm{T}}(\omega_0)\right|^2$ are defining Γ_1 such that we simply have

$$|\Gamma_1| = \frac{2}{A_0\left|Y'_{\mathrm{T}}(\omega_0)\right|}.$$

The ISF theory also predicts that other noise processes can be up converted or down converted by the remaining Γ_k Fourier coefficients of the ISF. In the next section we will discuss $1/f$ noise and develop a generalized Kurokawa theory that will enable us to effectively calculate, from circuit considerations, the coefficient Γ_0 reponsible for the $1/f$-noise up conversion. Similarly the contribution of the higher Fourier coefficients Γ_k can be accounted for in the generalized Kurokawa formalism if the dependence of the nonlinear device current upon the higher harmonics of the device voltage is included. For this purpose, the nonlinear device current $Y_{\mathrm{IN}}(A, \omega_0) \times A$ can be replaced by the multitone Volterra expansions introduced in the previous chapter (Section 6.2) on behavioral modeling. In the limit of weak harmonic noise, the PHD approximation can then be used to linearize the dependence of the nonlinear device current at the fundamental frequency upon the perturbating noise voltage δV_k and δV_k^* at the harmonic frequency $k\omega_0$ for a noise perturbation $e_k(t) = \mathrm{Re}[\delta V_k(t)\exp(ik\omega_0 t)]$ located *anywhere* inside the circuit,

$$I_{\mathrm{IN}}(A, \omega, \delta V_k) = Y_{\mathrm{IN}}(A, \omega, 0)\ A + \sum_{k=1}^{\infty}\left[Y^{(1)}_{\mathrm{IN},k}(A_0, \omega_0, 0)\delta V_k(t) + Y^{(2)}_{\mathrm{IN},k}(A_0, \omega_0, 0)\delta V_k^*(t)\right].$$

The resulting noise sensitivity coefficient $|\Gamma_k|$ for $k \geq 1$ is then

$$|\Gamma_k|^2 = \frac{4}{A_0^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^4} \times \left\{\left[\mathbf{Y}'_{\mathrm{T}}(\omega_0)\cdot\left(\mathbf{Y}^{(1)}_{\mathrm{IN},k} + \mathbf{Y}^{(2)}_{\mathrm{IN},k}\right)\right]^2 + \left[\mathbf{Y}'_{\mathrm{T}}(\omega_0)\times\left(\mathbf{Y}^{(2)}_{\mathrm{IN},k} - \mathbf{Y}^{(1)}_{\mathrm{IN},k}\right)\right]^2\right\}.$$

Note that in accordance with the PHD theory we have $Y^{(2)}_{\mathrm{IN},1} = 0$ for $k = 1$ since no down conversion is involved for noise at ω_0. In devices with dominant cyclostationary noise such as shot noise (white noise proportional to the instantaneous device current) the down conversion of the noise at higher RF harmonics can be expected to play an important role. But, once the Γ_k for all the white-noise processes e_k have been evaluated, it is sufficient to replace $4|e|^2/\left(A_0^2\left|Y'_{\mathrm{T}}(\omega_0)\right|^2\right) = |e|^2|\Gamma_1|^2$ by $\sum_{k=0}^{\infty}|e_k|^2|\Gamma_k|^2$ in the Kurokawa white-noise spectral densities derived above. The advantage of the generalized Kurokawa formalism remains that the correlation between amplitude and phase noise (AM-to-PM conversion) is also accounted for via the factor α^2/β^2 in Equation (7.18).

7.5 Impact of $1/f$ noise on an oscillator

In this section we proceed with the extension of the Kurokawa analysis to $1/f$ noise. As we have seen, the operation of an oscillator can be linearized around a stable operating point. On the basis of this fact the Kurokawa analysis can be extended to account for $1/f$ noise by accounting for the variation of any DC current (or voltage) biasing $[I_{\mathrm{N}}(t) = I_{\mathrm{N0}} + \delta I_{\mathrm{N}}(t)]$ [22]. A single DC bias dependence is assumed in this work for the sake of simplicity, but an arbitrary number of DC bias dependences could be accounted for if needed. Using the above assumption, the admittance of the nonlinear part $Y_{\mathrm{IN}}(A, \omega, I_{\mathrm{N}})$ can be represented as [3], defining $Y_{\mathrm{IN},0} = Y_{\mathrm{IN}}(A_0, \omega_0, I_{\mathrm{N0}})$,

$$Y_{\mathrm{IN}}(A, \omega, I_{\mathrm{N}}) = Y_{\mathrm{IN}}(A_0, \omega_0, I_{\mathrm{N0}}) + \frac{\partial Y_{\mathrm{IN},0}}{\partial A}\,\delta A + \frac{\partial Y_{\mathrm{IN},0}}{\partial \omega}\,\delta\omega + \frac{\partial Y_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}}\,\delta I_{\mathrm{N}}, \tag{7.24}$$

where the operating point (A_0, ω_0, I_{N0}) represents the steady-state operating point without noise and other perturbing signals (Figure 7.17).

The amplitude (δA) and phase (ϕ) deviation can be represented by the following differential equations given by [3]:

$$\left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\right|^2 \frac{1}{A_0}\frac{d\delta A(t)}{dt} + \beta\,\delta A(t) = \mathcal{B}\,\delta I_{\mathrm{N}}(t), \tag{7.25}$$

$$\left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\right|^2 \frac{d\phi(t)}{dt} + \alpha\,\delta A(t) = -\mathcal{A}\,\delta I_{\mathrm{N}}(t), \tag{7.26}$$

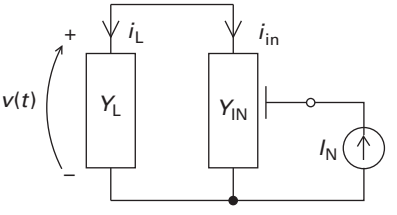


Figure 7.17 An admittance model of an oscillator with low-frequency modulation of the nonlinear device impedance at the fundamental frequency.

where $Y_{\mathrm{T},0} = Y_{\mathrm{T}}(A_0, \omega_0, I_{\mathrm{N}0})$ and where the two constants $\mathcal{A}$ and $\mathcal{B}$ are defined as

$$\mathcal{A} = \frac{\partial G_{\mathrm{T},0}}{\partial \omega}\frac{\partial G_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}} + \frac{\partial B_{\mathrm{T},0}}{\partial \omega}\frac{\partial B_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}},$$
$$\mathcal{B} = \frac{\partial G_{\mathrm{T},0}}{\partial \omega}\frac{\partial B_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}} - \frac{\partial B_{\mathrm{T},0}}{\partial \omega}\frac{\partial G_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}},$$

with

$$\frac{\partial Y_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}} = \frac{\partial G_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}} + j\,\frac{\partial B_{\mathrm{IN},0}}{\partial I_{\mathrm{N}}}.$$

The derivative terms $\partial Y_{\mathrm{IN},0}/\partial A$, $\partial Y_{\mathrm{IN},0}/\partial I_{\mathrm{N}}$, and $\partial Y_{\mathrm{T},0}/\partial \omega$ represent the variation of the admittances with perturbations, where G and B represent conductance and susceptance, respectively. The correlation factor α accounts for the correlation between the amplitude and phase deviations. $\delta I_{\mathrm{N}}(t)$ is the low-frequency $1/f$-noise current which will be defined by its power spectral density $S_{\delta I_{\mathrm{N}},1/f}(\Delta\omega) = S/|\Delta\omega|$ in the following subsection.

7.5.1 Derivation of $S_{a,1/f}(\Delta\omega)$

As was discussed in Chapter 4, the spectral density for a single trap is given by a Lorentzian spectrum

$$S_{\delta I_{\mathrm{N}},1\mathrm{trap}}(\Delta\omega) = \frac{2\lambda k}{\Delta\omega^2 + \lambda^2}. \tag{7.27}$$

A noise process with a $1/f$ distribution can then be obtained by a superposition of many of these Lorentzian distributions with time constants $\tau_{\mathrm{trap}}(y) = 1/\lambda = \tau_S \exp(\rho y)$ that are spatially varying with position y in the oxide (MOS) or wide-bandgap region (HFET):

$$\begin{aligned} S_{\delta I_{\mathrm{N}},1/f}(\Delta\omega) &= \int_0^{d_{\max}} S_{\delta I_{\mathrm{N}},1\mathrm{trap}}(\Delta\omega)dy \\ &= \frac{1}{\rho}\int_{\lambda_0}^{\lambda_1} S_{\delta I_{\mathrm{N}},1\mathrm{trap}}(\Delta\omega)\frac{(-d\lambda)}{\lambda} \\ &= \frac{2k}{\rho\,\Delta\omega}\left[\tan^{-1}\left(\frac{\lambda_0}{\Delta\omega}\right) - \tan^{-1}\left(\frac{\lambda_1}{\Delta\omega}\right)\right] \\ &\simeq \frac{k\pi}{\rho\,\Delta\omega} = \frac{S}{\Delta\omega} \qquad \text{for } \lambda_1 < \Delta\omega < \lambda_0, \end{aligned}$$

where $\lambda_0 = 1/\tau_S \simeq \infty$, $\lambda_1 = 1/\tau_{\mathrm{trap}}(d_{\max}) \simeq 0$ and $S = k\pi/\rho$. It is assumed that each of the traps captures electrons independently. This is usually justified since the traps are spatially distributed.

By taking the Fourier transform of (7.25) and (7.26), the following frequency-domain representations are obtained:

$$\left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\right|^2 \frac{1}{A_0}\,j\,\Delta\omega\,\delta A(\Delta\omega) + \beta\,\delta A(\Delta\omega) = \mathcal{B}\,\delta I_{\mathrm{N}}(\Delta\omega), \tag{7.28}$$

$$\left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\right|^2 j\,\Delta\omega\,\phi(\Delta\omega) + \alpha\,\delta A(\Delta\omega) = -\mathcal{A}\,\delta I_{\mathrm{N}}(\Delta\omega). \tag{7.29}$$

By solving (7.28), $\delta A(\Delta\omega)$ is calculated to be

$$\delta A(\Delta\omega) = \frac{\mathcal{B}\,\delta I_{\mathrm{N}}(\Delta\omega)}{\beta + \left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\right|^2 \frac{j\,\Delta\omega}{A_0}}. \tag{7.30}$$

Then the expression for the amplitude-noise spectral density for a single trap can be obtained using (7.27):

$$S_{\delta A,1\mathrm{trap}}(\Delta\omega) = \frac{2\mathcal{B}^2\lambda k}{(\Delta\omega^2+\lambda^2)\left(\beta^2 + \left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\right|^4 \frac{\Delta\omega^2}{A_0^2}\right)}, \tag{7.31}$$

where $0 < \Delta\omega < \omega_0$. This equation can be rewritten compactly as

$$S_{\delta A,1\mathrm{trap}}(\Delta\omega) = K_{\delta A}\left[\frac{1}{(\Delta\omega^2+k_2)(\Delta\omega^2+k_3)}\right], \tag{7.32}$$

where $K_{\delta A} = \left(2\mathcal{B}^2\lambda k A_0^2\right)/\left|\partial Y_{\mathrm{T},0}/\partial\omega\right|^4$, $k_2 = \lambda^2$, and $k_3 = \left(A_0^2\beta^2\right)/\left|\partial Y_{\mathrm{T},0}/\partial\omega\right|^4$.
By taking a summation over all traps in the limit of $\lambda_0 = \infty$ and $\lambda_1 = 0$, the amplitude-noise spectral density $S_{\delta A,1/f}(\Delta\omega)$ is derived to be

$$\begin{aligned} S_{\delta A,1/f}(\Delta\omega) &= \frac{1}{\rho}\int_{\lambda_0}^{\lambda_1} S_{\delta A,1\mathrm{trap}}(\Delta\omega)\left(\frac{-d\lambda}{\lambda}\right) \\ &= \frac{K_{\delta A}}{\rho\lambda}\frac{1}{\Delta\omega^2+k_3}\frac{1}{\Delta\omega}\left[\tan^{-1}\left(\frac{\lambda_0}{\Delta\omega}\right) - \tan^{-1}\left(\frac{\lambda_1}{\Delta\omega}\right)\right] \\ &\simeq \frac{S\mathcal{B}^2A_0^2}{\left|\partial Y_{\mathrm{T},0}/\partial\omega\right|^4}\frac{1}{\Delta\omega}\frac{1}{\Delta\omega^2+k_3} \qquad \text{for} \quad \lambda_1 < \Delta\omega < \lambda_2. \end{aligned} \tag{7.33}$$

Finally $S_{\delta A,1/f}(\Delta\omega)$ is normalized by A_0^2 to obtain the final analytic expression for the $1/f$ amplitude noise [23] [24]:

$$S_{a,1/f}(\Delta\omega) = \frac{S\mathcal{B}^2}{\left|\partial Y_{\mathrm{T},0}/\partial\omega\right|^4}\frac{1}{\Delta\omega}\frac{1}{(\Delta\omega^2+k_3)}. \tag{7.34}$$

7.5.2 Derivation of $S_{\phi,1/f}(\Delta\omega)$

To derive an expression for $S_{\phi,1\mathrm{trap}}(\Delta\omega)$, Equation (7.29) is solved in terms of $\phi(\omega)$, yielding

$$\phi(\Delta\omega) = -\frac{\mathcal{A}\,\delta I_{\mathrm{N}}(\Delta\omega) + \alpha\,\delta A(\Delta\omega)}{j\,\Delta\omega\left|\partial Y_{\mathrm{T},0}/\partial\omega\right|^2}. \tag{7.35}$$

Squaring Equation (7.35) yields the $1/f$ phase-noise spectral density for a single trap:

$$S_{\phi,1\mathrm{trap}}(\Delta\omega) = K_\phi\left[\frac{\Delta\omega^2+k_1}{\Delta\omega^2(\Delta\omega^2+k_2)(\Delta\omega^2+k_3)}\right], \tag{7.36}$$

where $K_\phi = 2\lambda k\kappa^2$, $\kappa = \mathcal{A}/\left|\partial Y_{\mathrm{T},0}/\partial\omega\right|^2$, $k_2 = \lambda^2$, and k_1 and k_3 given by

$$k_1 = \left[A_0^2(\beta\mathcal{A} + \alpha\mathcal{B})^2\right]/(\mathcal{A}^2|\partial Y_{\mathrm{T},0}/\partial\omega|^4),$$
$$k_3 = \left(A_0^2\beta^2\right)/|\partial Y_{\mathrm{T},0}/\partial\omega|^4.$$

By summing over all the traps as in Equation (7.33), the final analytic expression for the $1/f$ phase-noise spectral density $S_{\phi,1/f}(\Delta\omega)$ is obtained:

$$\begin{aligned} S_{\phi,1/f}(\Delta\omega) &= \frac{1}{\rho}\int_{\lambda_0}^{\lambda_1} S_{\phi,1\mathrm{trap}}(\Delta\omega)\frac{(-d\lambda)}{\lambda} \\ &= \frac{2k\kappa^2(\Delta\omega^2 + k_1)}{\rho\,\Delta\omega^2(\Delta\omega^2 + k_3)}\frac{1}{\Delta\omega}\left[\tan^{-1}\left(\frac{\lambda_0}{\Delta\omega}\right) - \tan^{-1}\left(\frac{\lambda_1}{\Delta\omega}\right)\right] \\ &= \frac{S\kappa^2(\Delta\omega^2 + k_1)}{\Delta\omega^3(\Delta\omega^2 + k_3)} \qquad \text{for } \lambda_1 < \Delta\omega < \lambda_0 \\ &\simeq \frac{S\kappa^2}{\Delta\omega^3}\frac{k_1}{k_3} = \frac{S\kappa^2}{\Delta\omega^3}\left(1 + \frac{\alpha\mathcal{B}}{\beta\mathcal{A}}\right)^2 \\ &\qquad \text{for } \lambda_1 < \Delta\omega < \min\left\{\sqrt{k_1}, \sqrt{k_3}, \lambda_0\right\}. \end{aligned} \tag{7.37}$$

The phase noise is seen to vary as $1/\Delta\omega^3$ for $1/f$ noise due to the frequency shaping resulting from the resonator. Here we have derived expressions for the amplitude and phase spectral density for $1/f$ noise that hold in the presence of correlation between amplitude and phase noise.

Expressions for the amplitude (AM) and phase (PM) components for the oscillator voltage spectral density $S_{V,1/f}(\Delta\omega)$ are given in Ref. [3]. Note that closed-form expressions are obtained for the autocorrelation $R_{\phi,1/f}(\tau)$ and $R_{\delta A,1/f}(\tau)$ (not reported). For the uncorrelated case ($k_3 = k_1$) the phase autocorrelation $R_{\phi,1/f}(\tau)$ reduces to

$$\begin{aligned} R_{\phi,1/f}(\tau) = \frac{k\kappa^2}{\rho}\Bigg[&\frac{e^{-\lambda_0|\tau|}}{2\lambda_0^2} - \frac{e^{-\lambda_1|\tau|}}{2\lambda_1^2} + \left(\frac{1}{\lambda_0} - \frac{e^{-\lambda_0|\tau|}}{2\lambda_0} - \frac{1}{\lambda_1} + \frac{e^{-\lambda_1|\tau|}}{2\lambda_1}\right)|\tau| \\ &+ \left(\mathrm{Ei}(-\lambda_1|\tau|) - \mathrm{Ei}(-\lambda_0|\tau|)\right)\frac{|\tau|^2}{2}\Bigg], \end{aligned} \tag{7.38}$$

with $\mathrm{Ei}(x)$ the exponential integral.

Substituting the analytic expressions for $R_{\phi,1/f}(\tau)$ into Equation (7.21) while optionally neglecting $R_{\delta A,1/f}(\tau)$ yields a closed-form expression for $R_{V,1/f}(\tau)$. The voltage noise density $S_{V,1/f}(\Delta\omega)$ can then be obtained by calculating numerically the Fourier transform of $R_{V,1/f}(\tau)$. Unlike the IEEE phase noise $S_{\phi,1/f}(\Delta\omega)/2$, which diverges at zero $\Delta\omega$, the voltage spectral density $S_{V,1/f}(\Delta\omega)$ (observable with a spectrum analyzer but not calculated by current commercial harmonic-balance simulators) is found to saturate at low offset frequencies (see the free-running oscillator phase noise in Figure 7.24 later for an experimental example and Ref. [3] for Kurokawa modeling results). This saturation is a requirement of power conservation, since the integration over frequency of the voltage spectral density $S_{V,1/f,\mathrm{ssb}}(\Delta\omega)$ yields the oscillator output power $\left(A_0^2/2\right)$. An approximate estimate of the ceiling voltage noise density $S_{V,1/f}(\mathrm{ceiling})$ and corner

frequency $\Delta f_{1/f}(\text{ceiling})$ can indeed be obtained from power conservation alone by assuming that the $1/\Delta f^3$ spectrum holds up to the ceiling corner frequency:

$$S_{V,1/f,\text{ssb}}(\text{ceiling}) = 0.6 \times \frac{A_0^3}{\sqrt{F}} \qquad \text{with} \quad F = \frac{A_0^2 S \kappa^2 k_1}{2\eta^2}$$

$$\Delta f_{1/f}(\text{ceiling}) = \frac{1}{2\pi}\left(\frac{F}{S_{V,1/f}(\text{ceiling})}\right)^{\frac{1}{3}}.$$

It is of interest to compare the Kurokawa result for $1/f$ noise with the result of the perturbation projection vector (PPV) [11] [12] for the nonlinear time-dependent techniques developed by Kaertner [10]. An analytic expression for $R_{\phi,1/f}(\Delta\omega)$ that is valid for small offset frequencies $\Delta\omega$ has recently been reported for the PPV method [13]. One can verify after numerical integration[4] that, for large measurement times, the PPV $R_{\phi,1/f}(\Delta\omega)$ (Equation (73) in Ref. [13]) yields essentially the same voltage spectral density $S_{V,1/f}(\Delta\omega)$ as that obtained from the Kurokawa $R_{\phi,1/f}(\Delta\omega)$ of Equation (7.38) independently of the strength S used for the $1/f$ noise. In particular, the same ceiling $S_{V,1/f}(\text{ceiling})$ is obtained from both numerical calculations for large measurement times. Furthermore, no nonlinear dependence on the $1/f$-noise strength S of the $1/\Delta\omega^3$ offset-frequency dependence of $S_{V,1/f}(\Delta\omega)$ is observed at high offset frequencies (away from saturation) for the PPV method, in agreement with the linear Kurokawa theory. Note that this close agreement in $S_{V,1/f}(\Delta\omega)$ for the two theories is obtained despite the fact that the autocorrelation $R_{\phi,1/f}$ in Ref. [13] is derived using the Keshner model for $1/f$ noise, whereas a sum of Ornstein–Uhlenbeck processes is used for modeling the $1/f$ noise in the above generalized Kurokawa derivation. The advantage of the generalized Kurokawa treatment remains that the correlation between amplitude and phase noise (AM to PM) is accounted for via the factor k_1/k_3 in the general expression for $R_{\phi,1/f}$ given by Equation (24) in Ref. [3] as well as in the IEEE phase noise $\mathcal{L} = S_{\phi,1/f}(\Delta\omega)$ given by Equation (7.37). The importance of the AM-to-PM conversion and the accuracy of the Kurokawa phase-noise results obtained for non-negligible correlation is further investigated in the next section.

Note that, in the presence of both white and $1/f$ noise, the correct approach for calculating the total voltage spectral density $S_{V,1/f}(\Delta\omega)$ from $R_{V,1/f}(\tau)$ is to first sum their respective phase and amplitude autocorrelations $R_{\phi,1/f}(\tau) + R_{\phi,\text{white}}(\tau)$ and $R_{\delta A,1/f}(\tau) + R_{\delta A,\text{white}}(\tau)$. Summing instead $S_{V,1/f}(\Delta\omega)$ and $S_{V,\text{white}}(\Delta\omega)$ can yield incorrect results for strong white noise. Indeed, the saturation of the voltage spectral density $S_{V,1/f}(\Delta\omega)$ for vanishing offset frequencies occurs only once, due to whichever noise process (white or $1/f$) induces it first (see Figure 12 in Ref. [3] for an example). No such issue arises for the IEEE phase noises $S_{\phi,i}(\Delta\omega)/2$ (and $S_{a,i}(\Delta\omega)$) of different uncorrelated processes i, which can just be added like $R_{\phi,i}(\tau)$ (and $R_{\delta A,i}(\tau)$) owing to the linearity of Equation (7.8).

Application of the white- and $1/f$-noise theory to the analysis of an *L–C* RFIC oscillator has been reported in Ref. [3]. A mode theory is used to calculate the effective noise

[4] Comparison to be reported elsewhere by Inwon Suh and the author.

source across the tank contributed by the four transistors involved in the two CMOS differential pairs. The contribution of the tail current is also accounted for.

7.5.3 Range of validity of the Kurokawa equations

Note that these analytic results for the Kurokawa theory were derived using a first-order perturbation of the nonlinear admittance Y_{L} and Y_{IN}. The validity of the first-order perturbation could be questioned for $1/f$ noise due to the diverging amplitude of the noise for low offset frequency. However, in practice the offset frequency assumed in the $1/f$ model is usually physically limited to a finite frequency range from $\Delta\omega_{\min}$ to $\Delta\omega_{\max}$. The validity of the Kurokawa approximation can then be verified by using a second-order perturbation (defining $Y_{\mathrm{T}} = Y_{\mathrm{L}} + Y_{\mathrm{IN}}$):

$$Y_{\mathrm{IN}}(A, \omega_0, I_{\mathrm{N0}}) \simeq Y_{\mathrm{IN},0} + \frac{\partial Y_{\mathrm{IN},0}}{\partial A}\,\delta A + \frac{1}{2}\frac{\partial^2 Y_{\mathrm{IN},0}}{\partial A^2}\,\delta A^2,$$
$$Y_{\mathrm{T}}(A_0, \omega, I_{\mathrm{N0}}) \simeq Y_{\mathrm{T},0} + \frac{\partial Y_{\mathrm{T},0}}{\partial \omega}\,\delta\omega + \frac{1}{2}\frac{\partial^2 Y_{\mathrm{T},0}}{\partial \omega^2}\,\delta\omega^2,$$

and verifying that the second-order terms are smaller than the first-order terms:

$$\left|\delta A^2(t)\right| \ll 4\left|\frac{\partial Y_{\mathrm{IN},0}}{\partial A} \Big/ \frac{\partial^2 Y_{\mathrm{IN},0}}{\partial A^2}\right|^2,$$
$$\left|\delta\omega^2(t)\right| \ll 4\left|\frac{\partial Y_{\mathrm{T},0}}{\partial \omega} \Big/ \frac{\partial^2 Y_{\mathrm{T},0}}{\partial \omega^2}\right|^2.$$

One can then determine whether these criteria are statistically met by using the average values $\overline{\left|\delta A^2\right|} = R_{\delta A}(0)$ and $\overline{\left|\delta\omega^2\right|} = R_{\delta\omega}(0)$, where their respective autocorrelations $R_{\delta A}(0)$ and $R_{\delta\omega}(0)$ at DC are obtained by integrating over frequency the spectral densities: $S_{\phi,1/f}(\Delta\omega)$ and $S_{\delta A,1/f}(\Delta\omega)$. These criteria are usually easily verified for $f_{\min}$ of 10^{-2} Hz or larger, for the $1/f$-noise strength S and correlation factor $\alpha\mathcal{B}/(\beta\mathcal{A})$ exhibited by common devices.

The circuit-based $1/f$-noise theory reported above can also be evaluated by comparing its results with those obtained from harmonic-balance simulators using the conversion-matrix methodology [15] [16]. For this purpose, a modified Van der Pol oscillator with nonlinear capacitances was developed to provide direct AM $1/f$-noise and PM $1/f$-noise generation for the uncorrelated case as well as AM-to-PM $1/f$-noise conversion for the correlated case [25].

Comparisons between the theoretical results and simulations obtained using the ADS harmonic-balance simulator are shown in Figures 7.18 and 7.19 for the amplitude and phase noise, respectively. Normalized correlation factors $\alpha\mathcal{B}/(\beta\mathcal{A})$ of -10, 0, and 10 dB are used for the three comparisons shown. A very close agreement is observed for the $1/f$ amplitude and $1/f$ phase noise between the analytic model and the circuit simulator for weak-to-medium AM-to-PM conversion. The phase noise calculated from $S_{V,1/f,\mathrm{ssb}}/\left(A_0^2/2\right)$ (old IEEE definition) is also shown in Figure 7.19.

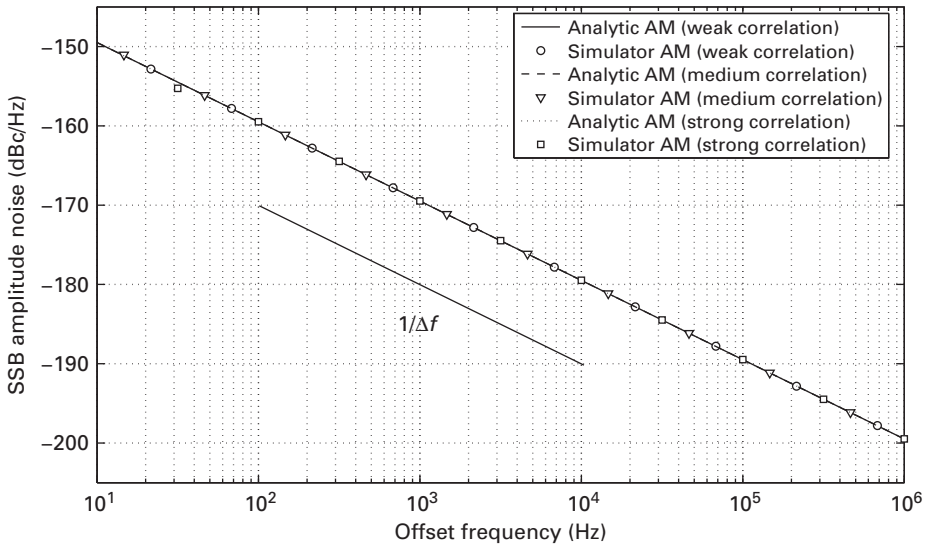


Figure 7.18 Comparison of harmonic-balance simulation (symbols) and modeling results (lines) obtained for amplitude noise using a correlation factor $\alpha\mathcal{B}/(\beta\mathcal{A})$ of -10, 0, and 10 dB. (Simulations performed by Inwon Suh at Ohio State University.)

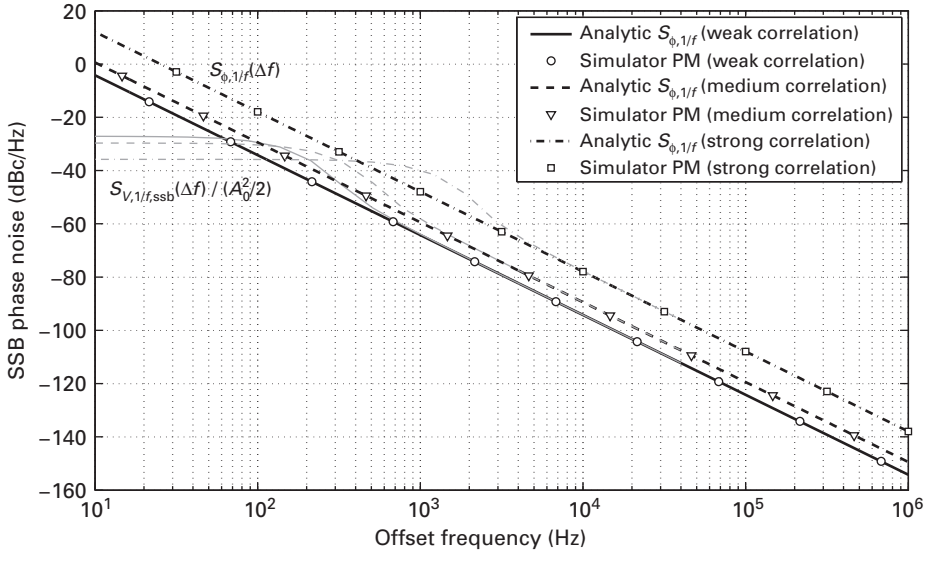


Figure 7.19 Comparison of harmonic-balance simulation (symbols) and modeling results (lines) obtained for phase noise using a correlation factor $\alpha\mathcal{B}/(\beta\mathcal{A})$ of -10, 0, and 10 dB. The phase noise calculated from $S_{V,1/f,\mathrm{ssb}}/\left(A_0^2/2\right)$ is shown using gray lines. (Simulations performed by Inwon Suh at Ohio State University.)

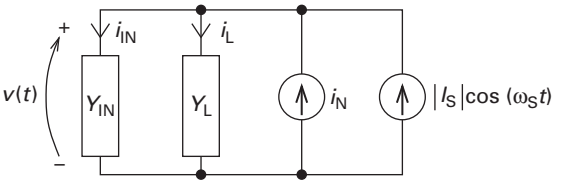


Figure 7.20 Equivalent circuit for the one-port injection-locked oscillator.

7.6 Injection locking and additive phase-noise measurements

The Kurokawa stability and noise theories reported in the previous sections were developed for a free-running one-port oscillator. The Kurokawa envelope analysis can also be applied to analyze injection-locked oscillators for both one-port [1] [2] and two-port [26] oscillator circuits. The equivalent circuit for the case of the one-port oscillator is shown in Figure 7.20. Under injection-locking conditions, the oscillator oscillates at the same frequency as that of the injected signal. Also, the phase of the oscillator is no longer slowly drifting due to the noise but is instead locked (with a constant offset) to the phase of the injected external signal.

7.6.1 Theory

The Kurokawa analysis of injection locking proceeds using Equations (7.5) and (7.6), where now the injected signal $i_N(t)$ is replaced by

$$i_N(t) \rightarrow |I_S|\cos(\omega_S t) + i_N(t),$$

with $|I_S|$ and ω_S the amplitude and frequency of the locking signal. Note that ω_S is selected to be close to the oscillating frequency ω_0 of the free-running oscillator. The oscillator noise current source $i_N(t)$ shown in Figure 7.20 is also included for completeness. The Kurokawa equations then reduce to

$$G_L(\omega_S) + G_{IN}[A(t), \omega_S] + G'_T(A_S, \omega_S)\frac{d\phi}{dt} + \frac{B'_T(A_S, \omega_S)}{A}\frac{dA}{dt} = \frac{|I_S|}{A_S}\cos\phi(t) + \frac{i_{N1}(t)}{A_S}, \tag{7.39}$$

$$B_L(\omega_S) + B_{IN}[A(t), \omega_S] + B'_T(A_S, \omega_S)\frac{d\phi}{dt} - \frac{G'_T(A_S, \omega_S)}{A}\frac{dA}{dt} = -\frac{|I_S|}{A_S}\sin\phi(t) - \frac{i_{N2}(t)}{A_S}. \tag{7.40}$$

Note that, under steady-state injection-locking conditions ($dA/dt = d\phi/dt = 0$, $\omega_i = \omega_S$, $A = A_S$, and $\phi = \phi_S$), the operating point is obtained from the following system of transcendental equations:

$$G_L(\omega_S) + G_{IN}(A_S, \omega_S) = \frac{|I_S|}{A_S}\cos\phi_S, \tag{7.41}$$

$$B_L(\omega_S) + B_{IN}(A_S, \omega_S) = -\frac{|I_S|}{A_S}\sin\phi_S, \tag{7.42}$$

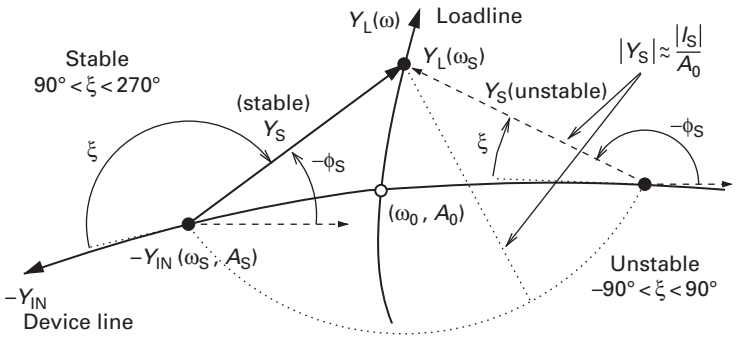


Figure 7.21 Kurokawa graphical analysis for injection-locking. The device line is assumed to be weakly frequency-dependent for simplicity.

where ϕ_S and A_S are the phase and amplitude, respectively, of the injection-locked oscillator voltage at the fundamental frequency. Given the input variables $|I_S|$ and ω_S, we can solve Equations (7.41) and (7.42) for the output variables ϕ_S and A_S.

A graphical analysis is constructed in Figure 7.21 from the vector addition $\mathbf{Y}_L = -\mathbf{Y}_{IN} + \mathbf{Y}_S$, where we introduced the injection vector $\mathbf{Y}_S = (|I_S|/A_S)[\cos(-\phi_S) \quad \sin(-\phi_S)]$ [1] with a phase $-\phi_S$ relative to the oscillator fundamental voltage taken as a reference. For a small injected signal I_S, the locked operating point (ω_S, A_S) is close to the free-running operating point (ω_0, A_0) and a perturbative analysis can be used for calculating the steady-state injection-locking operating point. Equations (7.41) and (7.42) then reduce to the following approximate linear system of equations:

$$G'_T(\omega_0)\delta\Omega_S + G'_{IN}(A_0, \omega_0)\delta A_S = \frac{|I_S|}{A_0}\cos\phi_S, \tag{7.43}$$

$$B'_T(\omega_0)\delta\Omega_S + B'_{IN}(A_0, \omega_0)\delta A_S = -\frac{|I_S|}{A_0}\sin\phi_S, \tag{7.44}$$

which is readily solved for $\delta A_S = A_S - A_0$ and $\delta\Omega_S = \omega_S - \omega_0$ in terms of $|I_S|$ and ϕ_S.

The stability and noise analysis for arbitrary injection power levels proceeds by assuming that the injection-locked operating point is slightly perturbed by fluctuations $\delta A(t)$ and $\delta\phi(t)$ of the oscillator voltage amplitude and phase at the fundamental frequency:

$$A(t) = A_S + \delta A(t) \quad \text{and} \quad \phi(t) = \phi_S + \delta\phi(t).$$

Subtracting the steady-state injection-locked solution from Equations (7.39) and (7.40) yields the following state equations for small time-varying amplitude $\delta A(t)$ and phase $\delta\phi(t)$ components:

$$\frac{\partial G_{IN}(A_S, \omega_S)}{\partial A}\delta A(t) + G'_T(A_S, \omega_S)\frac{d\delta\phi(t)}{dt} + \frac{B'_T(A_S, \omega_S)}{A_S}\frac{d\delta A(t)}{dt} = -\frac{|I_S|}{A_S}\sin\phi_S\,\delta\phi(t) + \frac{i_{N1}(t)}{A_S}, \tag{7.45}$$

$$\frac{\partial B_{\mathrm{IN}}(A_{\mathrm{S}},\omega_{\mathrm{S}})}{\partial A}\delta A(t)+B'_{\mathrm{T}}(A_{\mathrm{S}},\omega_{\mathrm{S}})\frac{d\delta\phi(t)}{dt}-\frac{G'_{\mathrm{T}}(A_{\mathrm{S}},\omega_{\mathrm{S}})}{A_{\mathrm{S}}}\frac{d\delta A(t)}{dt}$$
$$=-\frac{|I_{\mathrm{S}}|}{A_{\mathrm{S}}}\cos\phi_{\mathrm{S}}\,\delta\phi(t)-\frac{i_{\mathrm{N2}}(t)}{A_{\mathrm{S}}}. \tag{7.46}$$

The locking stability is then obtained by solving the characteristic equation $\left|Y'_{\mathrm{T,S}}\right|^2 s + bs + a = 0$ for the above system of equations with a and b as defined below. A stable lock is obtained when we have positive coefficients b and a:

$$b=\beta A_{\mathrm{S}}+\frac{|I_{\mathrm{S}}|}{A_{\mathrm{S}}}\left[G'_{\mathrm{T}}(A_{\mathrm{S}},\omega_{\mathrm{S}})\sin\phi_{\mathrm{S}}+B'_{\mathrm{T}}(A_{\mathrm{S}},\omega_{\mathrm{S}})\cos\phi_{\mathrm{S}}\right]>0, \tag{7.47}$$

$$a=|I_{\mathrm{S}}|\left[G'_{\mathrm{IN}}(A_{\mathrm{S}},\omega_{\mathrm{S}})\cos\phi_{\mathrm{S}}-B'_{\mathrm{IN}}(A_{\mathrm{S}},\omega_{\mathrm{S}})\sin\phi_{\mathrm{S}}\right]>0. \tag{7.48}$$

For low-power injected signals I_{S}, the first equation is usually satisfied, since as β is positive at the operating point $(A_0,\omega_0)\neq(A_{\mathrm{S}},\omega_{\mathrm{S}})$. The stability coefficient is then regulated by the second condition, which is graphically equivalent to the condition $-\mathbf{Y}'_{\mathrm{IN,S}}\cdot\mathbf{Y}_{\mathrm{S}}<0$ with $\mathbf{Y}'_{\mathrm{IN,S}}=\mathbf{Y}'_{\mathrm{IN}}(A_{\mathrm{S}},\omega_{\mathrm{S}})$. This stability condition can be written as $-\mathbf{Y}'_{\mathrm{IN,S}}\cdot\mathbf{Y}_{\mathrm{S}}=\left|\mathbf{Y}'_{\mathrm{IN,S}}\right||\mathbf{Y}_{\mathrm{S}}|\cos\xi<0$, where we define ξ as the angle from the device-line direction $-\mathbf{Y}'_{\mathrm{IN,S}}$ to the injection vector $\mathbf{Y}_{\mathrm{S}}$. It results that the angle ξ must be between 90° and 270° for a stable phase and frequency locking. Of the two possible solutions for $\mathbf{Y}_{\mathrm{S}}$ shown in Figure 7.21 for a given $|I_{\mathrm{S}}|$ and $\delta\omega_{\mathrm{S}}$, only the left one leads to a stable injection-locked oscillation.

For a weak injected signal $|I_{\mathrm{S}}|$, the locking phases $\phi_{\mathrm{S}\pm}$ at the edge of stability are obtained by setting $a=0$ in Equation (7.48). This is equivalent to the phase condition $\tan\phi_{\mathrm{S}\pm}=G'_{\mathrm{IN},0}/B'_{\mathrm{IN},0}$. Note that the oscillator phases $\phi_{\mathrm{S}\pm}$ correspond, respectively, to the angles $\xi_{\pm}$ of 90° or 270° shown in Figure 7.22.

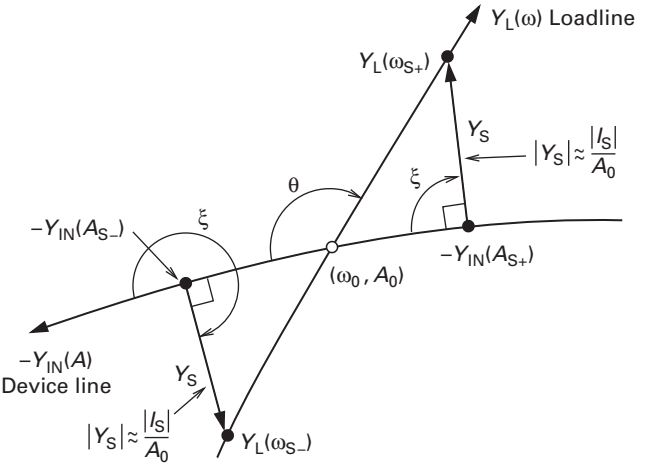


Figure 7.22 Graphical analysis for the locking-frequency range for a constant input power. The device line is assumed to be weakly frequency-dependent for simplicity.

Substituting the edge locking phases $\phi_{S\pm}$ into the solution of $\delta\Omega_S$ obtained from Equations (7.43) and (7.44) yields the locking frequency range $\Delta\Omega_S$:

$$\Delta\Omega_S = \delta\omega_{S+} - \delta\omega_{S-} = \frac{2|I_S|}{A_0} \times \frac{\left|Y'_{IN}(A_0,\omega_0)\right|}{|\beta|} = \frac{2|I_S|}{A_0\left|Y'_{T,0}\right|} \times \frac{1}{\sin\theta}, \tag{7.49}$$

with β the stability factor. The smaller the stability factor β, the wider the locking frequency range in the limit of small injection power levels. The locking-frequency range derived reduces to Adler's equation [27], if we set $B'_{IN,0} = 0$:

$$\Delta\Omega_S = \frac{2|I_S|}{A_0} \times \frac{1}{B'_T(A_0,\omega_0)} = \frac{|I_S|}{A_0} \times \frac{\omega_0}{QG_{T,0}},$$

with Q the oscillator quality factor given by $Q = \omega_0 B'_{T,0}/(2G_{T,0})$.

The output noise of the injection-locked oscillator [2] for a white-noise source $i_N(t)$ placed in shunt across the oscillator can also be analyzed using Equations (7.45) and (7.46). Limiting our analysis to the case of a weak correlation (α negligible), the additive phase noise is obtained by eliminating $d\delta A/dt$ from Equations (7.45) and (7.46) and using $\alpha = 0$. Note that the additive or residual phase noise denoted (+) is the phase noise contributed by the oscillator when the external injection signal is taken as the phase reference. It results that the additive phase-noise spectral density is

$$S_{\phi,\text{white}}^{(+),\text{locking}}(\Delta\omega) = \frac{2|e|^2}{A_0^2\left|Y'_{T,0}\right|^2} \times \frac{1}{\Delta\omega^2 + \Omega_C^2}, \tag{7.50}$$

where the corner frequency Ω_C is defined as

$$\Omega_C = \frac{|I_S|}{A_0} \frac{\left|B'_{T,0}\cos\phi_S + G'_{T,0}\sin\phi_S\right|}{\left|Y'_{T,0}\right|^2} = \frac{\mathbf{Y}_{S,0} \times \mathbf{Y}'_{T,0}}{\left|Y'_{T,0}\right|^2} = \frac{|I_S|\sin\chi}{A_0\left|Y'_{T,0}\right|},$$

with χ the anti-clockwise angle from $\mathbf{Y}_{S,0}$ to $\mathbf{Y}'_{T,0}$. It is to be noted that the corner frequency is proportional to the locking bandwidth [26] given by Equation (7.49).

It results from Equation (7.50) that for offset frequencies $\Delta\omega$ larger than Ω_C the additive phase-noise spectral density of the injection-locked oscillator is given by

$$S_{\phi,\text{white}}^{(+),\text{locking}}(\Delta\omega) \simeq \frac{2|e|^2}{A_0^2\left|Y'_{T,0}\right|^2\,\Delta\omega^2},$$

which is the white phase-noise spectral density (with $\alpha = 0$) already derived in Equation (7.18) in the absence of injection locking. For offset frequencies $\Delta\omega$ larger than Ω_C, the additive phase noise of the injection-locked oscillator has therefore reduced to the phase noise of the free-running oscillator

On the other hand, it results from Equation (7.50) that, for offset frequencies $\Delta\omega$ smaller than Ω_C, the additive phase-noise spectral density of the injection-locked oscillator reduces to a constant white noise given by

$$S_{\phi,\text{white}}^{(+),\text{locking}}(\Delta\omega=0)=\frac{2|e|^2|Y_{\text{T},0}|^2}{|I_{\text{S}}|^2\left|B'_{\text{T},0}\cos\phi_{\text{S}}+G'_{\text{T},0}\sin\phi_{\text{S}}\right|^2}=\frac{2|e|^2}{|I_{\text{S}}|^2}\frac{1}{\sin^2\chi}.$$

This residual additive white noise $S_{\phi,\text{white}}^{(+),\text{locking}}(\Delta\omega=0)$ is minimized for χ close to 90° and $|I_{\text{S}}|$ large compared with the white-noise spectral density $|e|^2$. For offset frequencies smaller than Ω_{S}, the residual additive white-noise spectral density $S_{\phi,\text{white}}^{(+),\text{locking}}(\Delta\omega=0)$ is smaller than the phase-noise spectral density of the free-running oscillator, which is proportional to $1/\Delta\omega^2$. Clearly the injection locking has prevented the generation of $1/\Delta\omega^2$ noise normally induced by the oscillator white noise $i_{\text{N}}(t)$. Using the model introduced in Section 7.5, one can verify that this result holds also for $1/f$ noise;[5] the injection lock prevents the generation of $1/\Delta\omega^3$ noise from the $1/\Delta\omega$ modulation noise of the device admittance Y_{IN}. Thus injection locking can be used to reduce the noise of a high-power oscillator by locking it to a less-noisy and lower-power oscillator.

An extension of this noise theory to two-port injection-locked oscillators has been reported in Ref. [26]. In the two-port injection-locked oscillator circuit, one port is used for the injection signal and the other port for the oscillator output. As in the case of $1/f$ noise, the nonlinear device impedance is perturbatively modulated by the injected external signal assuming low injected power levels. Closed-form expressions are derived for the phase noise, which exhibits the same trend as that for the one-port injection-locked oscillator [2]. Namely, the phase noise of the injection-locked oscillator follows that of the external source for small offset frequencies or that of the free-running oscillator for large offset frequencies. A relationship between the locking bandwidth and the corner frequency at which the phase noise switches from that of the external source to the free-running oscillator has also been established [26].

7.6.2 Experimental measurements

In the real-time active load-pull discussed in Section 7.3 a stable negative resistance was characterized. It is also possible to test an oscillator (unstable circuit) with an NVNA if the oscillator frequency is locked to an external RF signal source. In the case of a single-port oscillator, this can be achieved with the testbed shown in Figure 7.23. In this testbed a circulator is used for the signal injection. As a result the injection-locked oscillator behaves much like a nonlinear unstable amplifier between port 1 and port 2. A specific frequency range and power range is required at port 1 for locking of the oscillator. The Kurokawa stability analysis [1] for the one-port injection-locked oscillator presented in the previous section is directly applicable to the testbed shown in Figure 7.23.

The additive phase noise contributed by the locked oscillator can also be characterized using the injection-locking testbed equipped with a phase detector as shown in Figure 7.23. The merit of the additive phase-noise scheme is to remove the noise contributed by the signal source used for locking the oscillator. Experimental results

[5] Derivation to be reported elsewhere by Inwon Suh and the author.

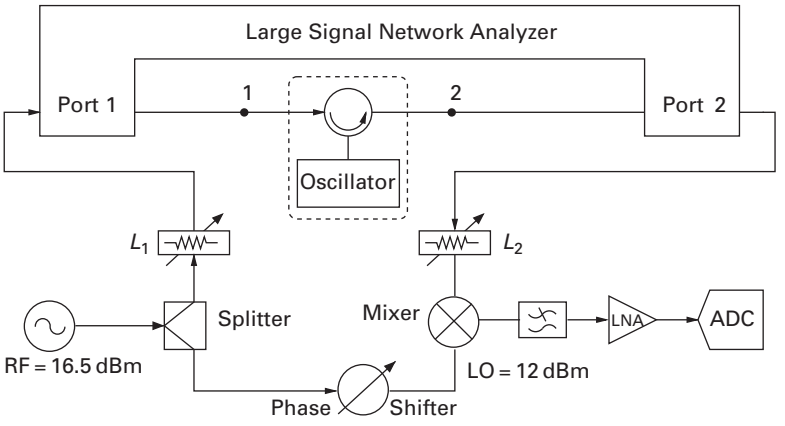


Figure 7.23 The LSNA testbed used for additive phase-noise measurements of an injection-locked oscillator.

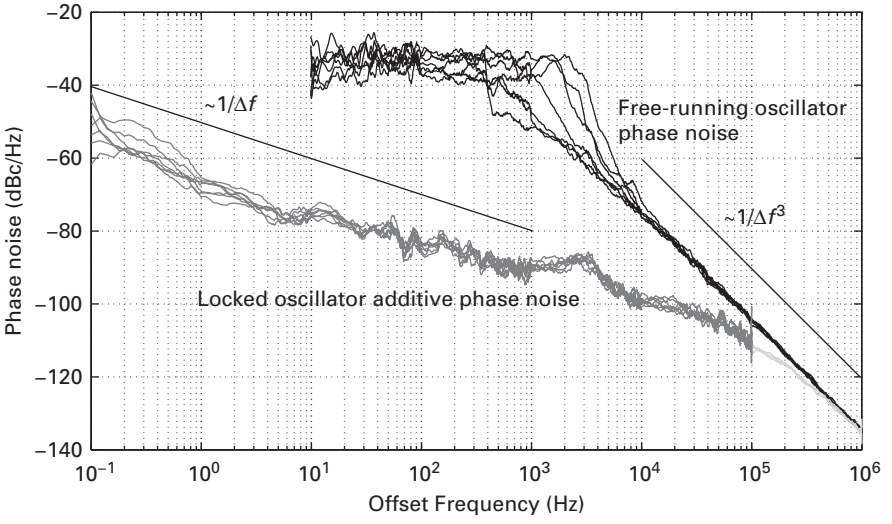


Figure 7.24 Phase noise for a free-running oscillator (black lines) and a locked oscillator (light gray lines) and additive phase noise for a locked oscillator (dark gray lines). (Measured by Inwon Suh at Ohio State University.)

are shown in Figure 7.24 for a pHEMT oscillator. The black lines represent successive measurements of the phase noise for the free-running oscillator. The dark gray lines represent additive phase-noise measurements (limited to offset frequencies below 0.1 MHz) for the noise contributed by the oscillator circuit under injection locking. The light gray lines are phase-noise measurements above 0.1 MHz of the injection-locked oscillator.

The free-running oscillator (black lines) is exhibiting phase noise with a $1/\Delta f^3$ dependence from 1 kHz to 1 MHz. The additive phase noise measured below 0.1 MHz (dark gray lines) is seen to have an approximate $1/\Delta f$ dependence. The additive

phase-noise curves (dark gray lines) of the locked oscillator are seen to intercept the phase-noise curves (black lines) of the free-running oscillator at the noise corner frequency ($\Omega_C/(2\pi)$ of 0.1 MHz. Given the relationship established in Ref. [26] between the noise corner frequency and the locking bandwidth, the phase-noise data can be interpreted as follows. Below 0.1 MHz, the offset frequency of the noise is below the locking bandwidth $\Delta\Omega_S/(2\pi)$ of the injection-locked oscillator, and the up conversion of the $1/\Delta f$ noise of the negative resistance to $1/\Delta f^3$ is rejected by the phase-locking mechanism. Conversely, above 0.1 MHz, the offset frequency of the noise is above the locking bandwidth $\Delta\Omega_S/(2\pi)$ of the injection-locked oscillator, and the phase-locking mechanism is no longer able to prevent the up conversion of the $1/\Delta f$ noise of the negative resistance to $1/\Delta f^3$ phase noise. Indeed, the comparison of the dark and light gray lines indicates that the phase noises measured for the free-running and locked oscillators exhibit the same magnitude and $1/\Delta f^3$ dependence above 0.1 MHz.

The testbed shown in Figure 7.23 can also be used to characterize a two-port oscillator as in Ref. [26] by connecting it between port 1 and port 2 in Figure 7.23. As for the one-port oscillator, the output voltage amplitude A_S, phase ϕ_S, and noise of the two-port oscillator can then be characterized as functions of both the frequency ω_S and the amplitude of the injected signal at port 1.

References

[1] K. Kurokawa, *Microwave Solid State Oscillator Circuits*, 2nd edition, Addison-Wesley, 1968.

[2] K. Kurokawa, "Noise in synchronized oscillators," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 16, No. 4, pp. 234–240, April 1968.

[3] J. Mukherjee, P. Roblin, and S. Bibyk, "An analytic circuit-based model for white and flicker phase noise in *LC* oscillators," *IEEE Transactions on Circuits and Systems*, Vol. 54, No. 7, pp. 1584–1598, July 2007.

[4] W. Wagner, "Oscillator design by device line measurement," *Microwave Journal*, Vol. 22, No. 2, p. 43, Feb. 1979.

[5] M. Vehovec, L. Houselander, and R. Spence, "On oscillator design for maximum power," *IEEE Transactions on Circuits and Systems*, Vol. 15, No. 3, pp. 281–283, Mar. 1968.

[6] V. M. T. Lam, P. C. L. Yip, and C. R. Poole, "Microwave oscillator design with power prediction," *Electronics Letters*, Vol. 27, No. 17, pp. 1574–1575, Sept. 1991.

[7] P. Berini, M. Desgagne, F. M. Ghannouchi, and R. G. Bosisio, "An experimental study of the effects of harmonic loading on microwave MESFET oscillators and amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 42, No. 6, pp. 943–950, June 1994.

[8] I. Suh, S. J. Doo, P. Roblin, X. Cui, Y. G. Kim, J. Strahler, M. Vanden Bossche, R. Rojas, and H. D. Park, "Negative input resistance and real-time active loadpull measurements of a 2.5 GHz oscillator using an LSNA," in *69th ARFTG Conference Digest, Honolulu, HI*, June 2007.

[9] I. Suh, P. Roblin, S. J. Doo, X. Cui, J. Strahler, and R. G. Rojas, "Measurement-based methodology to design harmonic loaded oscillators using real-time active load-pull," *IET Microwaves, Antennas & Propagation*, Vol. 5, No. 1, pp. 77–83, 2011.

[10] F. X. Kaertner, "Analysis of white and $f^{-\alpha}$ noise in oscillators," *International Journal of Circuit Theory and Applications*, Vol. 18, No. 5, pp. 485–519, Sept. 1990.

[11] J. Roychowdhury, A. Demir, and A. Mehrotra, "Phase noise in oscillators: a unifying theory and numerical methods for characterization," *IEEE Transactions on Circuits and Systems I*, Vol. 47, No. 4, pp. 655–674, May 2000.

[12] A. Demir, "Phase noise and timing jitter in oscillators with colored-noise sources," *IEEE Transactions on Circuits and Systems I*, Vol. 49, No. 12, pp. 1782–1791, Dec. 2002.

[13] S. Sancho, A. Suárez, J. Domínguez, and F. Ramírez, "Analysis of near-carrier phase-noise spectrum in free-running oscillators in the presence of white and colored noise sources," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 58, No. 3, pp. 587–601, Mar. 2010.

[14] A. Hajimiri and T. H. Lee, "A general theory of phase noise in electrical oscillators," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 2, pp. 179–194, 1998.

[15] P. Penfield, "Circuit theory of periodically driven nonlinear systems," *Proceedings of the IEEE*, Vol. 54, No. 2, pp. 266–280, Feb. 1966.

[16] V. Rizzoli, F. Mastri, and D. Masotti, "General noise analysis of nonlinear microwave circuits by the piecewise harmonic balance technique," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 42, No. 5, pp. 807–819, May 1994.

[17] A. S. Roy and C. C. Enz, "Analytical modeling of large-signal cyclo-stationary low-frequency noise with arbitrary periodic input," *IEEE Transactions on Electron Devices*, Vol. 54, No. 9, pp. 2537–2545, Sept. 2007.

[18] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd edition, McGraw Hill, 1984.

[19] E. S. Ferre-Pikal, J. R. Vig, J. C. Camparo, L. S. Cutler, L. Maleki, W. J. Riley, S. R. Stein, C. Thomas, F. L. Walls, and J. D. White, "Draft revision of IEEE STD 1139-1988 standard definitions of physical quantities for fundamental frequency theory and applications," in *Proceedings of the IEEE International Frequency Control Symposium*, pp. 338–357, 1997.

[20] D. B. Leeson, "A simple model of feedback oscillator noise spectrum," *Proceedings of the IEEE*, Vol. 54, No. 2, pp. 329–330, Feb. 1966.

[21] J. Bank, "A Harmonic-Oscillator Design Methodology Based on Describing Functions," Dissertation, Doctor of Philosophy, Chalmers University of Technology, 2006.

[22] H. Rohdin, Chung-Yi Su, and C. Stolte, "A study of the relation between device low-frequency noise and oscillator phase noise for GaAs MESFETs," in *IEEE Microwave Symposium Digest*, pp. 267–269, 1984.

[23] L. M. Nelson, C. W. Nelson, and F. L. Walls, "Relationship of AM to PM noise in selected RF oscillators," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, Vol. 41, No. 5, pp. 680–684, Sept. 1994.

[24] E. Rubiola, *Phase Noise and Frequency Stability in Oscillators*, Cambridge University Press, 2009.

[25] I. Suh, P. Roblin, and H. Jang, "Generalized Van Der Pol Oscillator for $1/f$ Noise," unpublished.

[26] F. Ramírez, M. Pontón, S. Sancho, and A. Suárez, "Phase-noise analysis of injection-locked oscillators and analog frequency dividers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 56, No. 2, pp. 393–407, Feb. 2008.

[27] R. Adler, "A study of locking phenomena in oscillators," *Proceedings of the IEEE*, Vol. 61, No. 10, pp. 1380–1385, Oct. 1973.

8 Design, modeling, and linearization of mixers, modulators, and demodulators[1]

This chapter is concerned with the system design, behavioral modeling, experimental characterization, balancing, and linearization of *I–Q* modulators and mixers. The K-parameters are first introduced for the purpose of linear modeling and balancing. The extraction of the K-matrix from LSNA measurements and a modulator–demodulator chain is discussed. The polyphase up-converter topology which permits the realization of linear single-sideband mixers is introduced next. The nonlinear modeling of the single-sideband mixer is then discussed. Finally the linearization of single-sideband mixers using the poly-harmonic predistortion technique is presented. These techniques offer some interesting options for the filterless implementation of wideband software-defined radio (SDR).

8.1 Vector characterization of an *I–Q* modulator

8.1.1 Balancing of an *I–Q* modulator

The balancing of *I–Q* modulators and demodulators is of critical importance in wireless systems, especially for modern multi-carrier techniques (OFDM) employing higher-order modulations such as QAM-64. This problem has been approached from the receiver side using adaptive equalization techniques using known preambles [1] [2]. Balancing is also of importance in multi-band predistortion linearization of the transmitter power amplifiers. As we shall see in Chapter 9 when introducing a frequency-selective predistortion linearization algorithm that differentially linearizes the signal in different bands [3], a balanced modulator is required in order for this linearization scheme to be effective.

An effective technique has been developed for balancing modulators [4]. This approach relies on the initial K-matrix modeling of *I–Q* modulators [5]. The extraction of the K-matrix of modulators using large-signal network analyzer measurements [6] [7] will be discussed in the next section.

[1] Research collaboration with Xi Yang, Dominique Chaillot, Suk Keun Myoung, and Shashank Mutha is gratefully acknowledged.

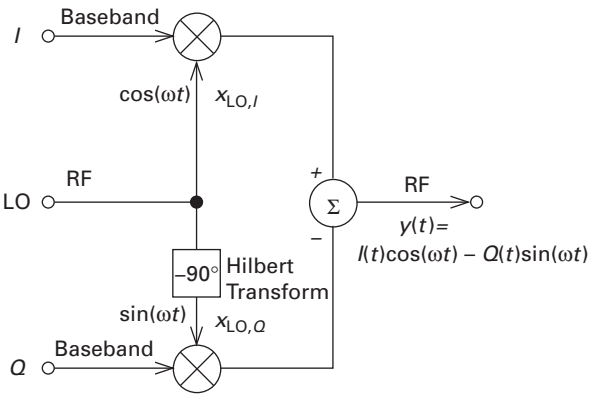


Figure 8.1 Ideal *I*–*Q* modulator used for up conversion.

8.1.2 K modeling

Consider the ideal *I*–*Q* modulator shown in Figure 8.1. The output of this *I*–*Q* modulator is

$$x_{\mathrm{RF}} = I\cos(\omega_0 t) - Q\sin(\omega_0 t).$$

A non-ideal *I*–*Q* modulator can contribute a gain and phase error, for example, due to imperfect in-phase and quadrature LO signals:

$$x_{\mathrm{LO},I} = \left(1 + \frac{\epsilon}{2}\right)\cos\left(\omega_0 t + \frac{\theta}{2}\right), \tag{8.1}$$

$$x_{\mathrm{LO},Q} = \left(1 - \frac{\epsilon}{2}\right)\sin\left(\omega_0 t - \frac{\theta}{2}\right), \tag{8.2}$$

where ϵ and θ are the amplitude and phase errors, respectively.

The resulting I_{RF} and Q_{RF} baseband components after *I*–*Q* modulation are then

$$\begin{aligned} x_{\mathrm{RF}} &= I x_{\mathrm{LO},I} - Q x_{\mathrm{LO},Q} \\ &= I_{\mathrm{RF}}\cos(\omega_0 t) - Q_{\mathrm{RF}}\sin(\omega_0 t), \end{aligned}$$

where we have

$$\begin{bmatrix} I_{\mathrm{RF}}(t) \\ Q_{\mathrm{RF}}(t) \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\begin{bmatrix} I(t) \\ Q(t) \end{bmatrix} = \mathbf{K}_{\mathrm{m}}\begin{bmatrix} I(t) \\ Q(t) \end{bmatrix},$$

with $\mathbf{K}_{\mathrm{m}}$ given by

$$\mathbf{K}_{\mathrm{m}} = \begin{bmatrix} (1+\epsilon/2)\cos(\theta/2) & (1-\epsilon/2)\sin(\theta/2) \\ (1+\epsilon/2)\sin(\theta/2) & (1-\epsilon/2)\cos(\theta/2) \end{bmatrix}. \tag{8.3}$$

The notation K is used since this is sometimes referred to as the K-matrix [5].

To measure the impact of the *I*–*Q* imbalance, consider for example the transmitted constellation of an OFDM signal (see Figure 8.2) which is distorted by the imbalances in the modulator and the demodulator (see Figure 8.3).

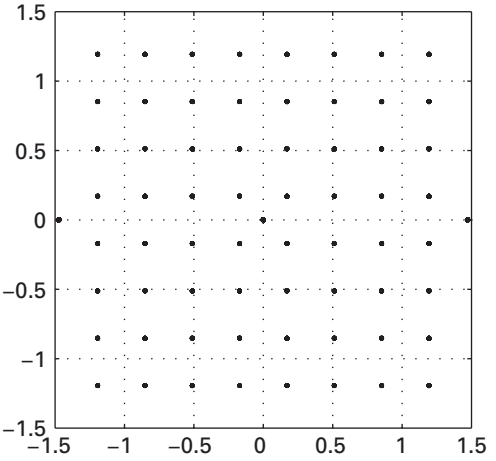


Figure 8.2 OFDM signal constellation.

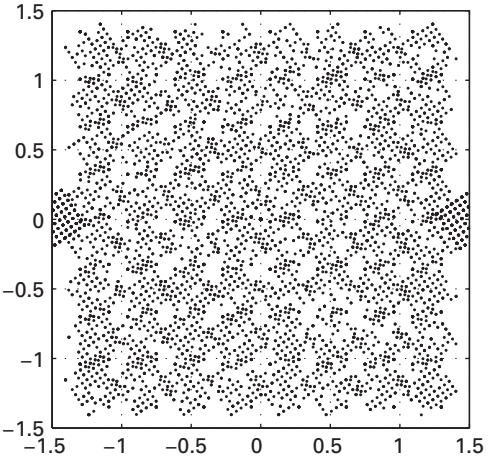


Figure 8.3 Effect of *I–Q* mismatch (7% amplitude error and 3° phase error) on OFDM signal constellation.

Note that the K model is a quasi-memoryless linear model as far the I and Q signals are concerned. In practice the K-matrix is found to be dependent on the spectral characteristics of $I(t)$ and $Q(t)$. If $I(t)$ and $Q(t)$ are band-limited in a band centered around ω_{m}, the matrix $\mathbf{K}_{\mathrm{m}}$ can be made dependent on the modulation frequency ω_{m}:

$$\begin{bmatrix} I_{\mathrm{RF}}(t) \\ Q_{\mathrm{RF}}(t) \end{bmatrix} = \begin{bmatrix} K_{11}(\omega_{\mathrm{m}}) & K_{12}(\omega_{\mathrm{m}}) \\ K_{21}(\omega_{\mathrm{m}}) & K_{22}(\omega_{\mathrm{m}}) \end{bmatrix} \begin{bmatrix} I(t) \\ Q(t) \end{bmatrix}.$$

This piecewise quasi-memoryless approximation finds application in the up conversion of multi-band signals.

To correct for the non-ideal effect introduced by a non-diagonal K-matrix, it is common practice to predistort the I and Q input signals [4]:

$$\begin{bmatrix} I'(t) \\ Q'(t) \end{bmatrix} = \begin{bmatrix} M_{11}(\omega_m) & M_{12}(\omega_m) \\ M_{21}(\omega_m) & M_{22}(\omega_m) \end{bmatrix} \begin{bmatrix} I(t) \\ Q(t) \end{bmatrix} = \mathbf{M} \begin{bmatrix} I(t) \\ Q(t) \end{bmatrix},$$

such that $\mathbf{M}_{PD} \times \mathbf{K} = \mathbf{U}$, with $\mathbf{U}$ the identity matrix. So $\mathbf{M}_{PD}$ is the inverse of $\mathbf{K}$:

$$\mathbf{M}_{PD} = \mathbf{K}^{-1} = \frac{1}{1-\epsilon^2/4} \frac{1}{\cos^2(\theta/2) - \sin^2(\theta/2)} \times \begin{bmatrix} (1-\epsilon/2)\cos(\theta/2) & -(1-\epsilon/2)\sin(\theta/2) \\ -(1+\epsilon/2)\sin(\theta/2) & (1+\epsilon/2)\cos(\theta/2) \end{bmatrix}. \quad (8.4)$$

8.1.3 *I–Q* modulator characterization with LSNA

An LSNA equipped with a hardware trigger can be used to characterize *I–Q* modulators. Either an input trigger or an output trigger can be accepted by the ADCs of the LSNA to start the measurement. The input trigger is favored because it enables one to lock the trigger to the modulation frequency. As shown in Figure 8.4, the I and Q signals are generated by a baseband generator that shares the same time base (10 MHz) as the LSNA. The clock of the baseband generator is locked to the time base via a PLL. The impedance of the baseband and RF generators is 50 Ω. The I and Q waveform can be measured with a scope as well. In such a case the trigger used to start the LSNA measurements can also be used to initiate the scope measurements.

As indicated in Figure 8.4 the LO signal is measured at port 1 and the RF modulator output at port 2. This setup permits us to acquire the phase both of the LO signal and of the RF signal.

As we have seen above, the simplest model we can use to represent the *I–Q* imbalance of an *I–Q* modulator is the linear K model. Simple I and Q excitations can then be used to evaluate K_{11}, K_{12}, K_{21}, and K_{22}. Consider the experiment shown in Figure 8.4, where we inject $I = \cos(\omega_m t)$ and $Q = \sin(\omega_m t)$ into the modulator to generate the upper sideband of amplitude M_2 and phase ϕ_2 at the frequency $\omega_2 = \omega_0 + \omega_m$. Owing to the modulator imbalance, a parasitic lower sideband tone of amplitude M_1 and phase ϕ_1 at the frequency $\omega_1 = \omega_0 - \omega_m$ is also generated.

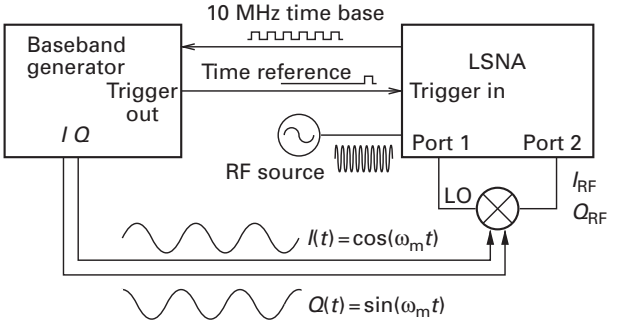


Figure 8.4 Setup for the characterization of an *I–Q* modulator for a given modulation frequency ω_m.

The K-matrix of the modulator can then be extracted from the phases and amplitudes ϕ_0, M_1, ϕ_1, M_2, and ϕ_2 measured by the LSNA (see Figure 8.5) using the following expression:

$$\begin{aligned}\mathbf{K}_{\mathrm{M}} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} &= M_1 \begin{bmatrix} \cos(\phi_1 - \phi_0) & \sin(\phi_1 - \phi_0) \\ \sin(\phi_1 - \phi_0) & -\cos(\phi_1 - \phi_0) \end{bmatrix} \\ &\quad + M_2 \begin{bmatrix} \cos(\phi_2 - \phi_0) & -\sin(\phi_2 - \phi_0) \\ \sin(\phi_2 - \phi_0) & \cos(\phi_2 - \phi_0) \end{bmatrix} \\ &= M_1 \mathbf{P}_1 + M_2 \mathbf{P}_2.\end{aligned}$$

Figure 8.6 shows the phases $\Phi_1 = \phi_1 - \phi_0$ and $\Phi_2 = \phi_2 - \phi_0$ obtained in 10 successive measurements. The standard deviations obtained for the trigger-in configuration

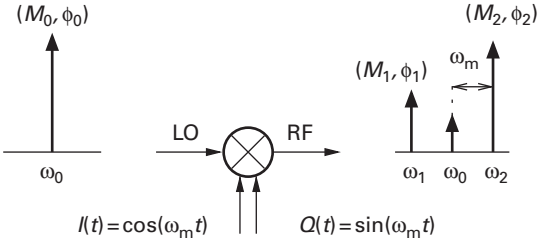


Figure 8.5 The I and Q signals are selected to generate the upper sideband (M_2,ϕ_1). The lower sideband (M_1,ϕ_1) arises from the imbalance of the *I–Q* modulator.

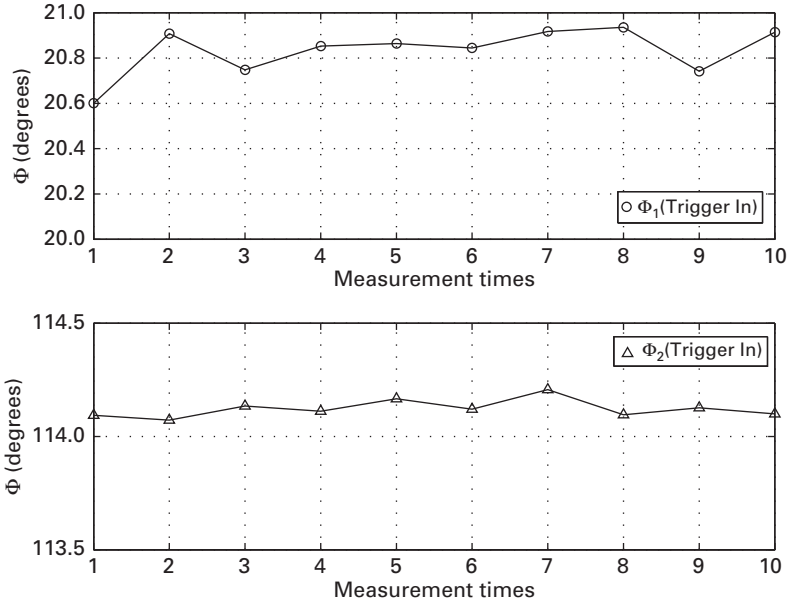


Figure 8.6 Variation of $\Phi_1 = \phi_1 - \phi_0$ and $\Phi_2 = \phi_2 - \phi_0$ in successive measurements when using the trigger-in configuration for center frequency 895 MHz and modulation frequency 1 MHz. (From [7] with permission, ©2005 IEEE.)

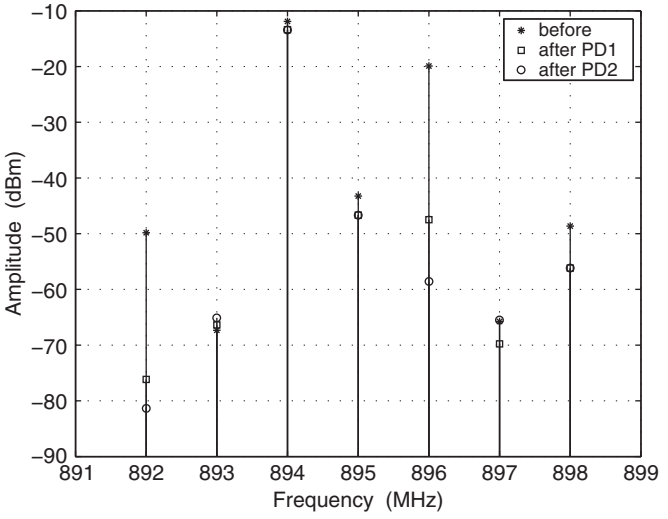


Figure 8.7 Cascading the *I–Q* correction K_{PD} and $K_{\mathrm{PD},2}$ yields a 46-dBc *I–Q* imbalance suppression at 896 MHz relative to 894 MHz for a lower-sideband excitation for the *I–Q* modulator tested. (From [7] with permission, ©2005 IEEE.)

are $\sigma_{\phi_1} = 0.1^\circ$ and $\sigma_{\phi_2} = 0.03^\circ$, respectively. The fluctuation of the phase is linked in part to the rise time of the trigger, which is around 2 ns in this experiment.

To balance the *I–Q* modulator the matrix $\mathbf{M}_{\mathrm{PD}}$ derived in Equation (8.4) can be used to predistort the *I–Q* data in the baseband generator. The results obtained are shown in Figure 8.7 using squares. Alternatively, the K-matrix measurement suggests the following inverse to predistort the *I–Q* data:

$$\mathbf{M}_{\mathrm{PD},2} = M_2 \mathbf{P}_2 (\mathbf{K}_{\mathrm{M}})^{-1}.$$

This *I–Q* conversion minimizes the change in phase by restoring the original group-phase shift of the *I–Q* modulator while removing its imbalance. Note that implementing a zero phase shift is unnecessary and could actually be undesirable. Indeed, the linear K model used is a simplification of the *I–Q* modulator response and it is best to use it only for relatively small amplitude and phase corrections. In fact, the best balancing results are obtained by cascading of the two predistortion methods. In other words, the predistortion using $\mathbf{M}_{\mathrm{PD2}}$ is applied after the modulator has been pre-balanced using $\mathbf{M}_{\mathrm{PD}}$ (PD1) in Equation (8.4). An *I–Q* imbalance suppression of 46 dBc and 43 dBc for the lower and upper sides, respectively, was achieved by cascading these two *I–Q* imbalance correction techniques as shown in Figure 8.7. Note that the balancing of the K-matrix needs be checked for both lower- and upper-sideband excitations using $I = \cos(\omega_{\mathrm{m}} t)$ and $Q = \pm\sin(\omega_{\mathrm{m}} t)$ if both excitations are to be used.

Better balancing obtained with a different *I–Q* modulator (see results in Section 8.3.2) reaches 60 to 70 dBc for a modulation frequency of 20 MHz. Note that in Figure 8.7 the nonlinear intermodulation products are of the same magnitude as the unwanted upper-sideband component. This indicates that the K model is an oversimplification of the modulator response and that nonlinear effects must be modeled as well.

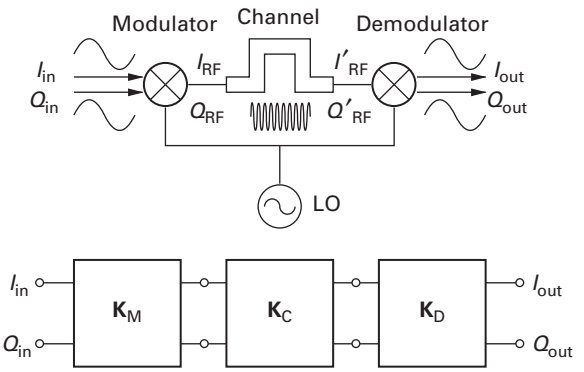


Figure 8.8 Cascaded *I–Q* modulator and *I–Q* demodulator with a shared LO. The channel used consists of various lossy or non-lossy delay lines.

8.1.4 K modeling of an *I–Q* modulator and an *I–Q* demodulator chain

This section presents a technique to balance an *I–Q* modulator and an *I–Q* demodulator cascaded together in a wireless transmitter receiver. The setup is shown in Figure 8.8. Since the *I–Q* modulator and *I–Q* demodulator are located in the same transmitter system, they can use the same LO signal for both up conversion and down conversion. Such a chain is typically used for the adaptive linearization of the power amplifier in wireless basestations.

It is possible to simultaneously extract the amplitude and phase errors in these two cascaded RF devices by inserting between them one or two calibrated delay lines in addition to the thru connection [8]. The actual number of delay lines needed depends on whether the delay-line losses can be neglected or not. An analytic equation can be derived for the extraction of the parameters of the K-matrices of the modulator and demodulator in terms of the measured K-matrices.

The signal path for the system shown in Figure 8.8 is as follows. The I and Q signals pass through the modulator, the channel, and the demodulator. Each of these modifies the *I–Q* signal according to its K-matrix.

We obtained the total K-matrix $\mathbf{K}_M$ of the modulator from the K-matrix $\mathbf{K}_m$ previously derived in Equation (8.3) by adding the modulator group delay ϕ_M and gain G_M:

$$\mathbf{K}_M = G_M \mathbf{K}_D(\phi_M) \times \begin{bmatrix} (1 + e_M/2)\cos(\theta_M/2) & (1 - e_M/2)\sin(\theta_M/2) \\ (1 + e_M/2)\sin(\theta_M/2) & (1 - e_M/2)\cos(\theta_M/2) \end{bmatrix},$$

with e_M the amplitude error and θ_M the phase error in the modulator.

The modulator K-matrix makes uses of the delay-line K-matrix $\mathbf{K}_g(\phi)$ defined as

$$\mathbf{K}_g(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix},$$

which transforms an input signal $x_{in}(t) = I\cos(\omega t) - Q\sin(\omega t)$ into an output signal $x_{out} = I\cos(\omega t + \phi) - Q\sin(\omega t + \phi)$.

The channel consists of switchable lossy delay lines. The K matrix $\mathbf{K}_{\mathrm{C},n}$ for the various delay lines used is therefore given by:

$$\mathbf{K}_{\mathrm{C},n} = G_{\mathrm{c},n}\mathbf{K}_{\mathrm{g}}(\phi_{\mathrm{c},n}),$$

where n is the index for the delay lines A, B, and C, $G_{\mathrm{c},n}$ the delay-line gains (usually smaller than unity), and $\phi_{\mathrm{c},n}$ the delay-line group phases.

The demodulator K-matrix $\mathbf{K}_{\mathrm{D}}$ is given by

$$\mathbf{K}_{\mathrm{D}} = G_{\mathrm{D}}\begin{bmatrix} (1+e_{\mathrm{D}}/2)\cos(\theta_{\mathrm{D}}/2) & (1+e_{\mathrm{D}}/2)\sin(\theta_{\mathrm{D}}/2) \\ (1-e_{\mathrm{D}}/2)\sin(\theta_{\mathrm{D}}/2) & (1-e_{\mathrm{D}}/2)\cos(\theta_{\mathrm{D}}/2) \end{bmatrix} \times \mathbf{K}_{\mathrm{g}}(\phi_{\mathrm{D}}),$$

with G_{D} the gain, ϕ_{D} the group delay phase, e_{D} the amplitude error, and θ_{D} the phase error in the demodulator. The I and Q signals are indeed obtained as follows:

$$I_{\mathrm{out}}(t) = \mathrm{LP}\big[2x'_{\mathrm{in}}(t) \times x'_{\mathrm{LO},I}(t)\big],$$
$$Q_{\mathrm{out}}(t) = \mathrm{LP}\Big[-2x'_{\mathrm{in}}(t) \times x'_{\mathrm{LO},Q}(t)\Big],$$

where LP indicates lowpass filtering. Note that $x_{\mathrm{in}} = I\cos(\omega t) - Q\sin(\omega t)$ in the input signal on the demodulator, and $x'_{\mathrm{in}}(t) = G_{\mathrm{D}}\left[I\cos(\omega t + \phi_{\mathrm{D}}) - Q\sin(\omega t + \phi_{\mathrm{D}})\right]$ is the input signal phase-shifted by the demodulator group delay phase ϕ_{D} and amplified by the demodulator gain G_{D}. The variables $x_{\mathrm{LO},I}$ and $x_{\mathrm{LO},Q}$ stand for the imperfect in-phase and quadrature LO signals (see Equations (8.1) and (8.2)) used by the demodulator, which features an amplitude error e_{D} and a phase error θ_{D}.

The following definitions are made for notational convenience:

$$\begin{aligned} G_{\mathrm{T},n} &= G_{\mathrm{M}} \times G_{\mathrm{c},n} \times G_{\mathrm{D}}, \\ \theta_{\mathrm{c},n} &= \phi_{\mathrm{M}} + \phi_{\mathrm{c},n} + \phi_{\mathrm{D}}, \\ \psi_{1,n} &= (\theta_{\mathrm{M}}/2 + \theta_{\mathrm{c},n} - \theta_{\mathrm{D}}/2), \\ \psi_{2,n} &= (\theta_{\mathrm{M}}/2 - \theta_{\mathrm{c},n} + \theta_{\mathrm{D}}/2), \\ \psi_{3,n} &= (\theta_{\mathrm{M}}/2 + \theta_{\mathrm{c},n} + \theta_{\mathrm{D}}/2), \\ \psi_{4,n} &= (\theta_{\mathrm{M}}/2 - \theta_{\mathrm{c},n} - \theta_{\mathrm{D}}/2). \end{aligned}$$

The system K-matrix for each delay n is then

$$\mathbf{K}_{\mathrm{system},n} = \mathbf{K}_{\mathrm{D}} \times \mathbf{K}_{\mathrm{C},n} \times \mathbf{K}_{\mathrm{M}} = \begin{bmatrix} M_{11,n} & M_{12,n} \\ M_{21,n} & M_{22,n} \end{bmatrix},$$

where $M_{11,n}$, $M_{12,n}$, $M_{21,n}$, and $M_{22,n}$ are given by

$$M_{11,n} = G_{\mathrm{T},n}\left(1+\frac{e_{\mathrm{D}}}{2}\right)\left(1+\frac{e_{\mathrm{M}}}{2}\right)\cos(\psi_{1,n}),$$
$$M_{12,n} = G_{\mathrm{T},n}\left(1+\frac{e_{\mathrm{D}}}{2}\right)\left(1-\frac{e_{\mathrm{M}}}{2}\right)\sin(\psi_{2,n}),$$

$$M_{21,n} = G_{\mathrm{T},n}\left(1 - \frac{e_{\mathrm{D}}}{2}\right)\left(1 + \frac{e_{\mathrm{M}}}{2}\right)\sin(\psi_{3,n}),$$
$$M_{22,n} = G_{\mathrm{T},n}\left(1 - \frac{e_{\mathrm{D}}}{2}\right)\left(1 - \frac{e_{\mathrm{M}}}{2}\right)\cos(\psi_{4,n}).$$

Three $\mathbf{K}_{\mathrm{system},n}$ are obtained from the measurement of the A, B, and C delay lines. First the phases $\psi_{1,\mathrm{A}}$ and $\psi_{3,\mathrm{A}}$ are extracted from the following system of equations:

$$\begin{aligned}\tan(\psi_{1,\mathrm{A}}) + P_{(11),(21),(\mathrm{B,A})}\cot(\psi_{3,\mathrm{A}}) &= \cot\left(\Delta\phi_{\mathrm{c},(\mathrm{B,A})}\right)\left(P_{(11),(21),(\mathrm{B,A})} - 1\right),\\ \tan(\psi_{1,\mathrm{A}}) + P_{(11),(21),(\mathrm{C,A})}\cot(\psi_{3,\mathrm{A}}) &= \cot\left(\Delta\phi_{\mathrm{c},(\mathrm{C,A})}\right)\left(P_{(11),(21),(\mathrm{C,A})} - 1\right).\end{aligned}$$

Next the phases $\psi_{2,\mathrm{A}}$ and $\psi_{4,\mathrm{A}}$ are extracted from the following system of equations:

$$\begin{aligned}\tan(\psi_{4,\mathrm{A}}) + P_{(22),(12),(\mathrm{B,A})}\cot(\psi_{2,\mathrm{A}}) &= -\cot\left(\Delta\phi_{\mathrm{c},(\mathrm{B,A})}\right)\left(P_{(22),(12),(\mathrm{B,A})} - 1\right),\\ \tan(\psi_{4,\mathrm{A}}) + P_{(22),(12),(\mathrm{C,A})}\cot(\psi_{2,\mathrm{A}}) &= -\cot\left(\Delta\phi_{\mathrm{c},(\mathrm{C,A})}\right)\left(P_{(22),(12),(\mathrm{C,A})} - 1\right).\end{aligned}$$

The following definitions are used in the above systems of equations:

$$\begin{aligned}\Delta\phi_{\mathrm{c},(m,n)} &= \phi_{\mathrm{c},m} - \phi_{\mathrm{c},n},\\ R_{(ij),(k,l),n} &= M_{(ij),n}/M_{(k,l),n},\\ P_{(ij),(k,l),(m,n)} &= R_{(ij),(k,l),m}/R_{(ij),(k,l),n}.\end{aligned}$$

The final amplitude and phase errors are then obtained from

$$\begin{aligned} e_{\mathrm{M}} &= 2\frac{S_1 - 1}{S_1 + 1} \quad \text{and} \quad e_{\mathrm{D}} = 2\frac{S_2 - 1}{S_2 + 1},\\ \phi_{\mathrm{M}} &= 2\left(\psi_{1,\mathrm{A}} + \psi_{2,\mathrm{A}}\right) \quad \text{and} \quad \phi_{\mathrm{D}} = 2\left(\psi_{3,\mathrm{A}} - \psi_{1,\mathrm{A}}\right),\\ S_1 &= R_{(11),(12),\mathrm{A}}\frac{\sin(\psi_{2,\mathrm{A}})}{\cos(\psi_{1,\mathrm{A}})} \quad \text{and} \quad S_2 = R_{(11),(21),\mathrm{A}}\frac{\sin(\psi_{3,\mathrm{A}})}{\cos(\psi_{1,\mathrm{A}})}.\end{aligned}$$

Given the indeterminacy of the phase extraction using an arctangent function, four solutions are actually obtained and the unphysical ones need to be eliminated. For the case in which the delay-line losses are negligible, a single delay line beside a thru is required and the solution reduces to

$$\begin{aligned}\psi_{1,\mathrm{A}} &= \tan^{-1}\left\{\left[\cos(\theta_{\mathrm{delay}}) - M_{11,\mathrm{B}}/M_{11,\mathrm{A}}\right]/\sin(\theta_{\mathrm{delay}})\right\},\\ \psi_{2,\mathrm{A}} &= \cot^{-1}\left\{\left[\cos(\theta_{\mathrm{delay}}) - M_{12,\mathrm{B}}/M_{12,\mathrm{A}}\right]/\sin(\theta_{\mathrm{delay}})\right\},\\ \psi_{3,\mathrm{A}} &= \cot^{-1}\left\{\left[-\cos(\theta_{\mathrm{delay}}) + M_{21,\mathrm{B}}/M_{21,\mathrm{A}}\right]/\sin(\theta_{\mathrm{delay}})\right\},\\ \psi_{4,\mathrm{A}} &= \tan^{-1}\left\{\left[-\cos(\theta_{\mathrm{delay}}) + M_{22,\mathrm{B}}/M_{22,\mathrm{A}}\right]/\sin(\theta_{\mathrm{delay}})\right\},\\ \theta_{\mathrm{D}} &= \psi_{1,\mathrm{A}} + \psi_{2,\mathrm{A}},\\ \theta_{\mathrm{M}} &= \frac{1}{2}\left(-\psi_{1,\mathrm{A}} + \psi_{2,\mathrm{A}} + \psi_{3,\mathrm{A}} - \psi_{4,\mathrm{A}}\right),\end{aligned}$$

$$e_{\mathrm{M}} = 2\frac{M_{11,\mathrm{A}}/M_{12,\mathrm{A}} - \cos(\psi_{1,\mathrm{A}})/\sin(\psi_{2,\mathrm{A}})}{M_{11,\mathrm{A}}/M_{12,\mathrm{A}} + \cos(\psi_{1,\mathrm{A}})/\sin(\psi_{2,\mathrm{A}})},$$
$$e_{\mathrm{D}} = 2\frac{M_{11,\mathrm{A}}/M_{21,\mathrm{A}} - \cos(\psi_{1,\mathrm{A}})/\sin(\psi_{3,\mathrm{A}})}{M_{11,\mathrm{A}}/M_{21,\mathrm{A}} + \cos(\psi_{1,\mathrm{A}})/\sin(\psi_{3,\mathrm{A}})},$$

with θ_{delay} defined as $\theta_{\mathrm{c,B}} = \theta_{\mathrm{c,A}} + \theta_{\mathrm{delay}}$.

An alternative approach relies on two different LO frequencies for the modulator and demodulator as shown in Figure 8.9 and the use of Fourier analysis. In such a case it is no longer necessary to rely on extra delay lines to characterize the modulator–demodulator chain. However, the modulator and demodulator K-matrices must be weakly frequency-dependent (memoryless). This can be approximated by using a small enough offset LO frequency at the cost of a larger Fourier analysis.

The use of a different LO frequency ω_{dem} for the demodulator from the modulator frequency ω_{mod} effectively changes the time reference for the I and Q definition such that the insertion of an extra K-matrix is required. Consider a signal $x(t)$ present at the demodulator input:

$$x(t) = I\cos(\omega_{\mathrm{mod}}t) - Q\sin(\omega_{\mathrm{mod}}t) = I'\cos(\omega_{\mathrm{dem}}t) - Q'\sin(\omega_{\mathrm{dem}}t).$$

The following K-matrix $\mathbf{K}_{\Delta\omega}(t)$ relating the input I and Q to the output I' and Q' is obtained:

$$\mathbf{K}_{\Delta\omega}(t) = \begin{bmatrix} \cos(\Delta\omega\, t) & -\sin(\Delta\omega\, t) \\ \sin(\Delta\omega\, t) & \cos(\Delta\omega\, t) \end{bmatrix},$$

using the offset LO frequency $\Delta\omega = \omega_{\mathrm{mod}} - \omega_{\mathrm{dem}}$. The K-matrix $\mathbf{K}_{\Delta\omega}(t)$ associated with a change of LO is therefore the same as that of the delay line $\mathbf{K}_{\mathrm{g}}(\phi)$ except that now the phase shift $\phi = \Delta\omega\, t$ is time-dependent.

The total K-matrix for the system is then

$$\mathbf{K}_{\mathrm{system,2LO}} = \mathbf{K}_{\mathrm{D}} \times \mathbf{K}_{\mathrm{C}} \times \mathbf{K}_{\Delta\omega}(t) \times \mathbf{K}_{\mathrm{M}} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

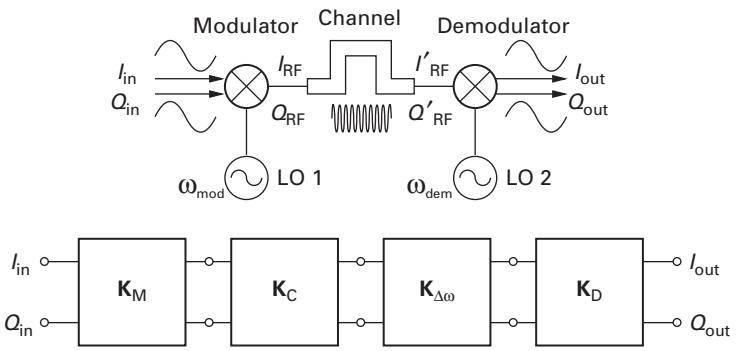


Figure 8.9 Cascaded I–Q modulator and I–Q demodulator with two different LO frequencies.

where M_{11}, M_{12}, M_{21}, and M_{22} are given by

$$
\begin{aligned}
M_{11} &= G_{\mathrm{T}}\left(1+\frac{e_{\mathrm{D}}}{2}\right)\left(1+\frac{e_{\mathrm{M}}}{2}\right)\cos(\Delta\omega t+\psi_1),\\
M_{12} &= -G_{\mathrm{T}}\left(1+\frac{e_{\mathrm{D}}}{2}\right)\left(1-\frac{e_{\mathrm{M}}}{2}\right)\sin(\Delta\omega t-\psi_2),\\
M_{21} &= G_{\mathrm{T}}\left(1-\frac{e_{\mathrm{D}}}{2}\right)\left(1+\frac{e_{\mathrm{M}}}{2}\right)\sin(\Delta\omega t+\psi_3),\\
M_{22} &= G_{\mathrm{T}}\left(1-\frac{e_{\mathrm{D}}}{2}\right)\left(1-\frac{e_{\mathrm{M}}}{2}\right)\cos(\Delta\omega t-\psi_4),
\end{aligned}
$$

where the gain and phase parameters are still the same but defined for a single channel:

$$
\begin{aligned}
G_{\mathrm{T},n} &= G_{\mathrm{M}}\times G_{\mathrm{c}}\times G_{\mathrm{D}},\\
\theta_{\mathrm{c}}(t) &= \phi_{\mathrm{M}}+\phi_{\mathrm{c}}+\phi_{\mathrm{D}},\\
\psi_1 &= (\theta_{\mathrm{M}}/2+\theta_{\mathrm{c}}-\theta_{\mathrm{D}}/2),\\
\psi_2 &= (\theta_{\mathrm{M}}/2-\theta_{\mathrm{c}}+\theta_{\mathrm{D}}/2),\\
\psi_3 &= (\theta_{\mathrm{M}}/2+\theta_{\mathrm{c}}+\theta_{\mathrm{D}}/2),\\
\psi_4 &= (\theta_{\mathrm{M}}/2-\theta_{\mathrm{c}}-\theta_{\mathrm{D}}/2).
\end{aligned}
$$

Let us assume that a single tone ω_{m} is used to probe the system response such that

$$
I = I_0\cos(\omega_{\mathrm{m}}t) \qquad \text{and} \qquad Q = Q_0\sin(\omega_{\mathrm{m}}t).
$$

The output I and Q signals are then

$$
\begin{aligned}
I_{\mathrm{out}}(t, I_0, Q_0) = {} & I_0 2G_{\mathrm{T}}\left(1+\frac{e_{\mathrm{D}}}{2}\right)\left(1+\frac{e_{\mathrm{M}}}{2}\right)\\
&\times\{\cos[(\omega_{\mathrm{m}}+\Delta\omega)t+\psi_1]+\cos[(\omega_{\mathrm{m}}-\Delta\omega)t-\psi_1]\}\\
&+ Q_0 2G_{\mathrm{T}}\left(1+\frac{e_{\mathrm{D}}}{2}\right)\left(1-\frac{e_{\mathrm{M}}}{2}\right)\\
&\times\{\cos[(\omega_{\mathrm{m}}+\Delta\omega)t-\psi_2]-\cos[(\omega_{\mathrm{m}}-\Delta\omega)t+\psi_2]\},\\
Q_{\mathrm{out}}(t, I_0, Q_0) = {} & I_0 2G_{\mathrm{T}}\left(1-\frac{e_{\mathrm{D}}}{2}\right)\left(1+\frac{e_{\mathrm{M}}}{2}\right)\\
&\times\{\sin[(\omega_{\mathrm{m}}+\Delta\omega)t+\psi_3]-\sin[(\omega_{\mathrm{m}}-\Delta\omega)t-\psi_3]\}\\
&+ Q_0 2G_{\mathrm{T}}\left(1-\frac{e_{\mathrm{D}}}{2}\right)\left(1-\frac{e_{\mathrm{M}}}{2}\right)\\
&\times\{\sin[(\omega_{\mathrm{m}}+\Delta\omega)t-\psi_4]+\sin[(\omega_{\mathrm{m}}-\Delta\omega)t+\psi_4]\}.
\end{aligned}
$$

We can then perform a Fourier transform (an FFT in a DSP implementation):

$$
\begin{aligned}
\mathcal{I}_{\mathrm{out}}(\omega, I_0, Q_0) &= F\left[I_{\mathrm{out}}(t, I_0, Q_0)\right],\\
\mathcal{Q}_{\mathrm{out}}(\omega, I_0, Q_0) &= F\left[Q_{\mathrm{out}}(t, I_0, Q_0)\right].
\end{aligned}
$$

If only the input $I(t)$ is excited, $I_0 = 1$ and $Q_0 = 0$, the following phases can be recovered:

$$\psi_1 = \angle \mathcal{I}_{\text{out}}(\omega, 1, 0) \qquad \text{and} \qquad \psi_3 = \angle \mathcal{Q}_{\text{out}}(\omega, 1, 0) + \frac{\pi}{2}.$$

If only the input $Q(t)$ is excited, $I_0 = 0$ and $Q_0 = 1$, the following phases can be recovered:

$$\psi_2 = -\angle \mathcal{I}_{\text{out}}(\omega, 0, 1) \qquad \text{and} \qquad \psi_4 = -\angle \mathcal{Q}_{\text{out}}(\omega, 0, 1) - \frac{\pi}{2}.$$

The modulator and demodulator phase errors are then given by

$$\phi_{\text{M}} = \psi_1 + \psi_2 \qquad \text{and} \qquad \phi_{\text{D}} = \psi_3 - \psi_1.$$

The modulator and demodulator amplitude errors are then given by

$$e_{\text{M}} = 2\left[\frac{|\mathcal{I}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 1, 0)|}{|\mathcal{I}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 0, 1)|} - 1\right]\left[\frac{|\mathcal{I}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 1, 0)|}{|\mathcal{I}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 0, 1)|} + 1\right]^{-1},$$

$$e_{\text{D}} = 2\left[\frac{|\mathcal{I}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 1, 0)|}{|\mathcal{Q}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 1, 0)|} - 1\right]\left[\frac{|\mathcal{I}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 1, 0)|}{|\mathcal{Q}_{\text{out}}(\omega_{\text{m}} + \Delta\omega, 1, 0)|} + 1\right]^{-1}.$$

This two-LO scheme can be beneficially integrated with the frequency-selective linearization algorithms to be presented for single-sideband mixers and multi-band power amplifiers in this chapter and in the next chapter, respectively. Indeed, the receiver bandwidth can be realized with a small bandwidth in these linearization schemes, but it then requires a frequency-agile demodulation LO signal ω_{dem} relative to the modulation LO frequency ω_{mod}.

8.2 Polyphase multi-path technique

Our discussion so far has been limited to the linear modulator. Nonlinear issues have become among the most challenging issues in the design of broadband multi-carrier and multi-band RF transmitters. Indeed, broadband multi-carrier modulation schemes exhibit high envelope fluctuation, which induces a strong nonlinear response in RF mixers and RF amplifiers, leading to spectral regrowth, inband distortion, and interference between adjacent channels. Various linearization strategies are then employed to reduce the nonlinearity or suppress the distortion products generated, while not degrading the power efficiency of the transmitter.

Filtering is one of the most commonly used methods to remove the unwanted distortion products. The design of filters providing the desired selectivity and insertion loss at RF frequencies is not without its challenges. Furthermore, different filters are required according to the different standards, creating a bottleneck for developing wireless transceiver terminals that work with multiple bands and multiple standards. In addition, filters cannot be used to remove inband distortion [9].

The polyphase multi-path technique has been proposed [9] [10] as an effective architecture for removing harmonics and intermodulation products. A well-known example

is the balanced circuit, a two-path polyphase circuit that is able to cancel out the even-order harmonics. By increasing the number of paths in the circuit, more distortion products can be canceled out. An N-path system will suppress $N-1$ spurious bands out of each group of N adjacent spurious bands. The method could facilitate the development of filterless software radios: one wideband integrated up converter with no dedicated external filters [10]. However, this technique has its own limitations. Mismatches between each path and the next will degrade the linearization performance and must be dealt with separately. Finally, the fundamental in-band intermodulation products, which are often the key issue in RF transmitters, cannot be removed by using a polyphase multi-path architecture without canceling out the desired signal. Having introduced the ideal multi-path mixer, we will then proceed in a subsequent section with the development of a behavioral model [11] to predict the nonlinear response of practical multi-path mixers. Armed with this realistic model, we shall then introduce a poly-harmonic predistortion linearization technique [11] to reduce the remaining uncanceled baseband harmonics and intermodulation products generated in practical polyphase multi-path circuits. Further discussions regarding the in-band linearization part will be postponed until Chapter 9.

8.2.1 Nonlinear behavior

To introduce the polyphase multi-path technique, we shall first consider a memoryless weakly nonlinear baseband circuit. Such a system can be modeled using the following Taylor series expansion:

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + a_3 x^3(t) + a_4 x^4(t) + a_5 x^5(t) + \cdots$$

When this system is excited by a two-tone baseband signal $x(t) = \cos(\omega_1 t) + \cos(\omega_2 t)$, the output then consists of the desired baseband sinusoidal signals with fundamental frequencies ω_1 and ω_2 plus undesired harmonics $n\omega_1$ and $n\omega_2$ and intermodulation products $k\omega_1 + m\omega_2$, with n, k, and m all integers.

8.2.2 Polyphase multi-path technique

A technique called polyphase multi-path linearization [9] and [10] can be used to cancel out distortion products. As is shown in Figure 8.10, the input signal is split into N paths. For each path, a first phase shifter is used to generate a sequence of signals with equally distributed phases, which are fed to N identical nonlinear devices, modeled by Equation (8.5). A second phase shifter is used to cancel out the first phase shift for the fundamental, and the N output signals are then combined. It results that, for the linear response of the device, the output signals of the N paths are in phase and add constructively, whereas for the nonlinear responses their phases are equally distributed and the output signals cancel each other out. In this way, the desired signal will be amplified/up converted while the unwanted distortions will be suppressed over a wide bandwidth, provided that a sufficiently large number of paths with equally distributed phase is selected.

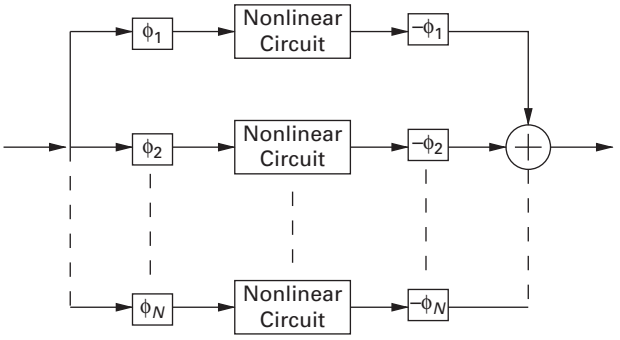


Figure 8.10 A polyphase multi-path circuit [9].

Assume the input signal $x(t) = A\cos(\omega t)$ is applied to the circuit. After the first phase shifter, the signal on each path is $x_i(t) = A\cos(\omega t + \phi_i)$ with $\phi_i = i2\pi/N$, where $i = 0, 1, \ldots, N-1$. The desired output of the nonlinear circuit will be $a_1 A\cos(\omega t + \phi_i)$. The second- and third-order nonlinear terms of Equation (8.5) will give

$$\begin{aligned} a_2 x^2(t) &= a_2(A\cos(\omega t + \phi_i))^2 \\ &= \frac{1}{2}a_2 A^2(1 + \cos(2\omega t + 2\phi_i)), \end{aligned} \tag{8.5}$$

$$\begin{aligned} a_3 x^3(t) &= a_3(A\cos(\omega t + \phi_i))^3 \\ &= \frac{1}{4}a_3 A^3(3\cos(\omega t + \phi_i) + \cos(3\omega t + 3\phi_i)). \end{aligned} \tag{8.6}$$

From Equations (8.5) and (8.6), we see that the nonlinear terms have the same effect on the frequency and the phase. One can infer that in general the phase of the nth harmonic of the nonlinear circuit will be $n\phi_i$.

After the second phase shifter, the phase of the nth harmonic before the adder becomes $(n-1)\phi_i = (n-1)i2\pi/N$. For the desired fundamental signal, $n = 1$, or for harmonics $n = pN + 1$ with $p = 1, 2, 3, \ldots$, the phase will reduce to 0 or a multiple of 2π in each path, and thus the signals will add constructively. For harmonics with n not equal to $pN + 1$, the phase in each path remains uniformly distributed over 2π such that these harmonics are suppressed when combined.

In this chapter, we focus on the application of the polyphase multi-path linearization to nonlinear mixers. Figure 8.11 shows a four-path polyphase circuit, where the second phase shifter is implemented by an ideal mixer to realize the wideband phase and frequency shifts [9] [12]. The nonlinear behavior of the mixers at baseband is represented by nonlinear circuits modeled by Equation (8.5) preceding the ideal linear mixers.

The input signal is divided into four paths with equally distributed phase (0°, 90°, 180°, 270°). Since the phase rotation for the kth harmonic is k times the input phase, the fundamental, the second harmonic, and the third harmonic will have phase shifts of (0°, 90°, 180°, 270°), (0°, 180°, 0°, 180°), and (0°, 270°, 180°, 90°), respectively, at the output of the nonlinear circuit. With the ideal mixer functioning as both a

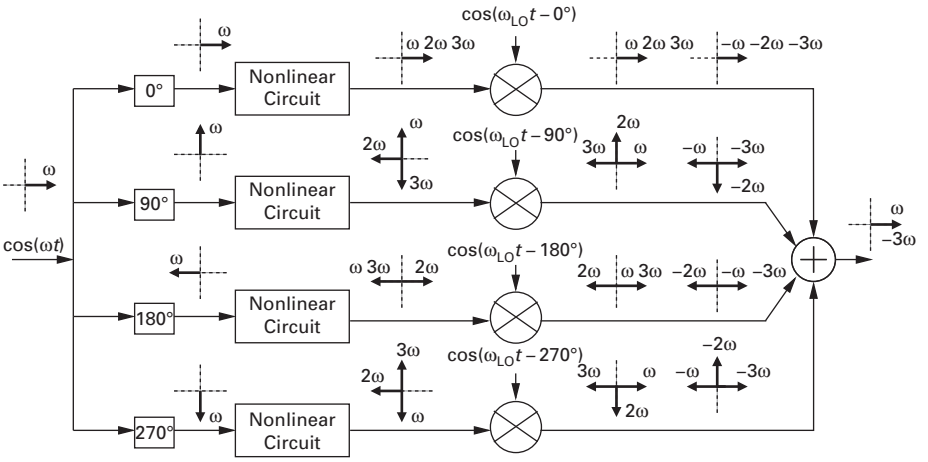


Figure 8.11 A four-path polyphase circuit using a mixer as a phase shifter.

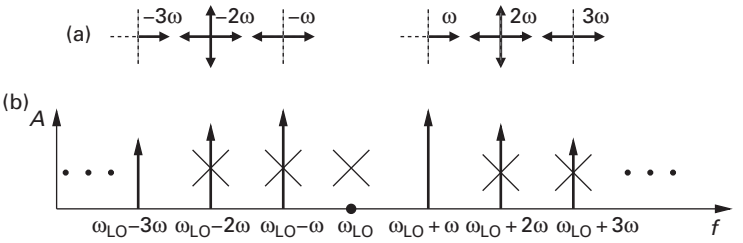


Figure 8.12 Output of the four-path polyphase circuit, (a) phasor diagrams and (b) output spectrum.

frequency shifter and a phase shifter, the harmonics of the nonlinear circuits will appear on both sides of the LO frequency due to the sum and difference frequencies resulting from mixing of the input and LO signals.

Figures 8.11 and 8.12 show the phases of the $\pm\omega$, $\pm 2\omega$, and $\pm 3\omega$ frequency components in each path at different stages, the phasor diagrams, and the final spectrum resulting at the output of the four-path polyphase mixer.

For the upper sideband (USB), the desired signals $\cos[(\omega_{\mathrm{LO}} + \omega)t]$ have the same phase at the end of each path and add constructively. However, for the second and third harmonics at $\omega_{\mathrm{LO}} + 2\omega$ and $\omega_{\mathrm{LO}} + 3\omega$, the phase of each path is equally distributed over 2π and results in signal cancellation when the four paths are added together. Using similar analysis, the fifth harmonic is found to have aligned phases, making it the first uncanceled harmonic in the USB.

For the lower sideband (LSB), the image of the fundamental at $\omega_{\mathrm{LO}} - \omega$ and that of the second harmonic at $\omega_{\mathrm{LO}} - 2\omega$ are similarly canceled out, because of their equally distributed phase over 2π. However, the third harmonic component at $\omega_{\mathrm{LO}} - 3\omega$ is not canceled out because the phases are the same at the end of each path such that the N-path output signals add constructively.

For an N-path polyphase circuit, with equally distributed phase, $\phi_i = i2\pi/N$ with $i = (0, 1, 2, \ldots, N-1)$, the uncanceled harmonic will be for $n = pN + 1$ ($p = \pm 1, \pm 2, \pm 3, \ldots$) [9]. If the input signal is a multitone signal, the intermodulation products at $k\omega_1 + m\omega_2$ with $k + m = n$ will be canceled out if the nth harmonic is canceled out. However, note that the fundamental in-band intermodulation products cannot be canceled out unless the desired signal is also canceled out. Indeed, the results derived above hold for the modulated case when the amplitude and phase are time-dependent in the limit of a memoryless system.

Figure 8.13 shows MATLAB simulation results for a two-tone excitation $x(t) = \cos(\omega_1 t) + \cos(\omega_2 t)$. The nonlinear circuit includes nonlinearities up to the seventh order. The output (top) of the mixer without polyphase multi-path techniques features up to seventh-order harmonic tones of the input frequency ω_1 and ω_2 on both sides

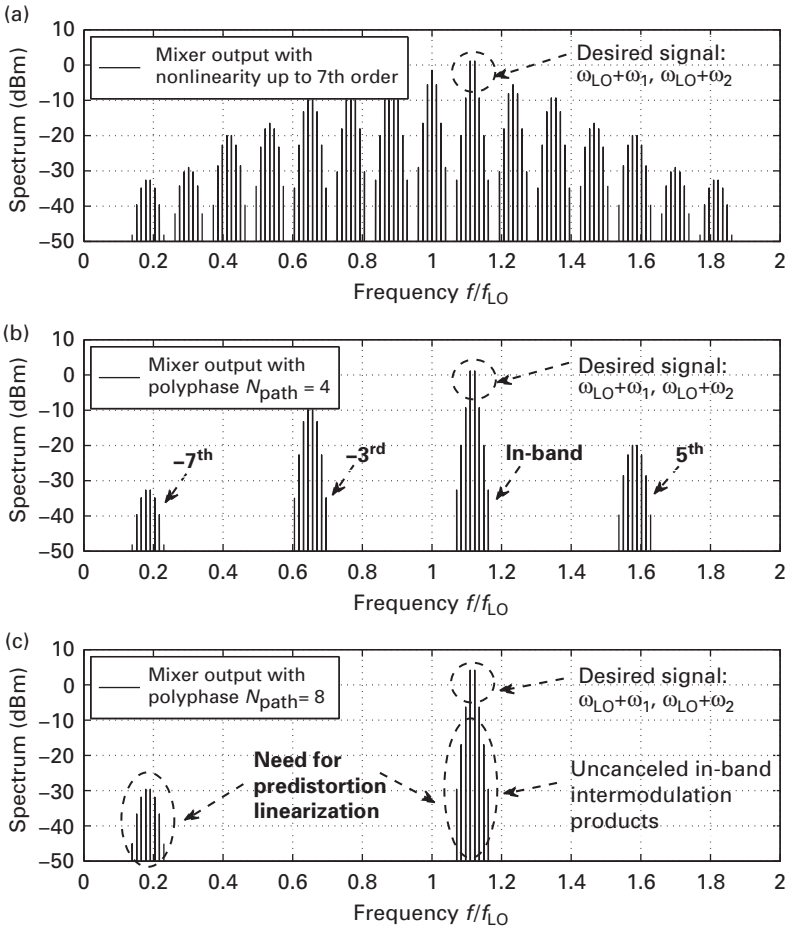


Figure 8.13 Simulated spectra of (a) a single, (b) a four-path, and (c) an eight-path polyphase mixer circuit in response to two-tone excitation.

of the LO. Each harmonic is surrounded by intermodulation products according to $k\omega_1 + m\omega_2$.

The four-path polyphase circuit preserves the desired output $\cos(\omega_{LO} + \omega_1)t$ and $\cos[(\omega_{LO} + \omega_2)t]$. The uncanceled harmonics are the -7th, -3rd, and 5th harmonics (the negative signs indicate that the harmonics appear at the LSB). The uncanceled intermodulation products are those bands associated with uncanceled harmonics, as well as those within the desired output signal band (in-band). The eight-path polyphase circuit results in additional distortion cancellation, leaving the -7th harmonic band uncanceled in addition to the in-band intermodulation products.

Note that we have focused on the frequency bands around the LO. The presence of harmonics of the LO signal will generate additional components. Some of them will also be suppressed by the polyphase architecture [9]. However, the LO harmonics can usually be easily removed using a lowpass filter if the transmitter bandwidth is smaller than the LO frequency. Thus we shall not be concerned here with these higher LO harmonic terms.

In summary, the polyphase multi-path technique is able to reduce the harmonics and intermodulation products generated by a memoryless weakly nonlinear circuit excited with a multi-carrier signal with wide envelope fluctuations. The larger the number of paths, the more harmonics and intermodulation products are canceled out. However, increasing the number of paths of the polyphase circuit increases the circuit complexity and cost. Memory effects and mismatches between each path and the next will degrade the ideal performance, yielding imperfect cancellation of undesired tones/bands. Furthermore, the intermodulation products within the desired output band will not be canceled out by the polyphase technique. Thus the polyphase multi-path technique will benefit from being supplemented by a linearization technique accounting for memory effects and compensating for mismatches to remove the remaining uncanceled harmonics and intermodulation products. The key to developing such a linearization scheme is to first develop a behavioral model for the polyphase circuits. This is pursued in the next section for a single-sideband mixer.

8.3 Poly-harmonic modeling of a single-sideband modulator

8.3.1 Theory

A common strategy used in microwave basestations is to rely on a digital IF for up converting a baseband signal to RF. One of the motivations is that the LO leakage and image band can then be easily suppressed since the RF band, the LO frequency, and the image band are widely separated. The baseband I–Q signals $I(t)$ and $Q(t)$ are first used at the digital level to modulate the IF tone ω_{IF}:

$$a_{IF} = I(t)\cos(\omega_{IF}t) - Q(t)\sin(\omega_{IF}t).$$

An analog single-sideband mixer is then used to shift the frequency from IF to RF. Such an ideal single-sideband mixer can be realized (or modeled) with an I–Q modulator:

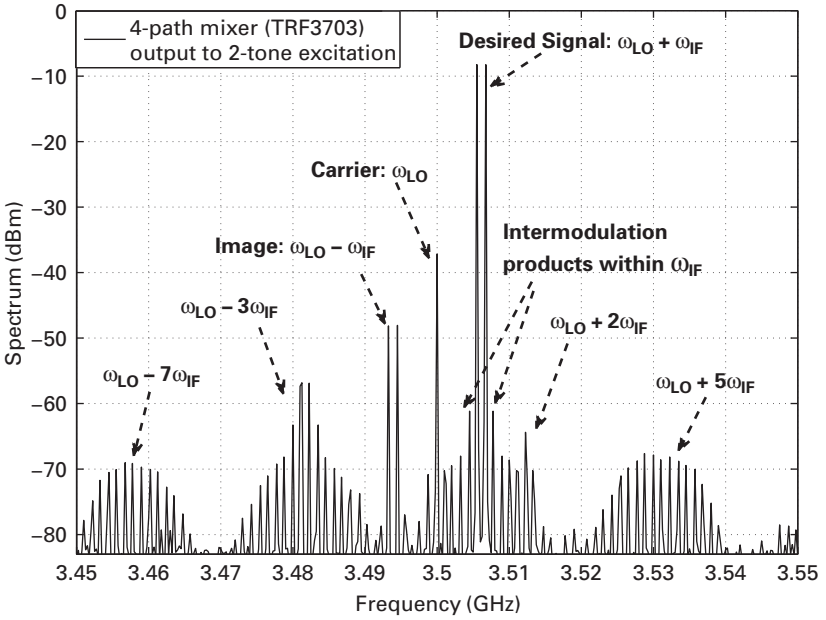


Figure 8.14 Output spectrum for a single-sideband mixer in response to a two-tone excitation in the strongly nonlinear regime. (Measured by Xi Yang at Ohio State University.)

$$b_{\mathrm{RF}} = I_{\mathrm{RF}}(t)\cos(\omega_{\mathrm{LO}}t) - Q_{\mathrm{RF}}(t)\sin(\omega_{\mathrm{LO}}t),$$

if we select $I_{\mathrm{RF}}(t) = a_{\mathrm{IF}}(t)$ and $Q_{\mathrm{RF}}(t) = \widehat{a_{\mathrm{IF}}(t)}$ for generating the upper sideband $\omega_{\mathrm{LO}} + \omega_{\mathrm{IF}}$ or if we select $I_{\mathrm{RF}}(t) = a_{\mathrm{IF}}(t)$ and $Q_{\mathrm{RF}}(t) = -\widehat{a_{\mathrm{IF}}(t)}$ for generating the lower sideband $\omega_{\mathrm{LO}} - \omega_{\mathrm{IF}}$. We shall assume that the upper sideband is the desired one.

In practice the analog *I–Q* modulator is not perfectly balanced and the lower-sideband image will also be detected, as we discussed at the begining of this chapter. Under large-signal excitations, additional spurious signals will also be generated. To illustrate these effects, the measured output spectrum from a single-sideband mixer implemented with a differential *I–Q* modulator for a very strong two-tone[2] input excitation is shown in Figure 8.14. Beside the desired signal around $\omega_{\mathrm{LO}} + \omega_{\mathrm{IF}}$ additions, spurious bands surrounding the desired band are observed at $\omega_{\mathrm{LO}} - 7\omega_{\mathrm{IF}}$, $\omega_{\mathrm{LO}} - 3\omega_{\mathrm{IF}}$, $\omega_{\mathrm{LO}} + 1\omega_{\mathrm{IF}}$ (the desired band), and $\omega_{\mathrm{LO}} + 5\omega_{\mathrm{IF}}$. A spacing of $4\omega_{\mathrm{IF}}$ is observed, in agreement with the polyphase theory, since, owing to the differential nature of its circuit, the single-sideband mixer is a polyphase circuit of order $N = 4$ (four paths). But, due to mismatches existing in the circuit, the image band at $\omega_{\mathrm{LO}} - \omega_{\mathrm{IF}}$, the LO leakage at ω_{LO}, and even a spurious band at $\omega_{\mathrm{LO}} + 2\omega_{\mathrm{IF}}$ are also observed.

Another feature to be noted is that the spurious and desired bands generated according to the polyphase topology, $\omega_{\mathrm{LO}} - 7\omega_{\mathrm{IF}}$, $\omega_{\mathrm{LO}} - 3\omega_{\mathrm{IF}}$, $\omega_{\mathrm{LO}} - \omega_{\mathrm{IF}}$, and $\omega_{\mathrm{LO}} + 5\omega_{\mathrm{IF}}$, all exhibit very dense spectral regrowth when this strongly nonlinear regime of operation

[2] The input I and Q excitations actually consist in this case of a single tone, but this situation is referred to as two-tone excitation because of the lower and upper sidebands generated at IF.

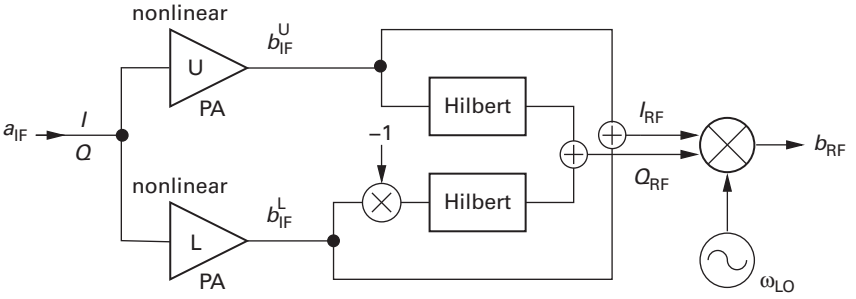


Figure 8.15 Behavioral model used for modeling the single-sideband mixer around the fundamental frequency.

is reached. Next we shall develop a behavioral model for such an up converter that can predict a similar behavior, although operation in this strongly nonlinear mode is always avoided.

In our effort to develop a behavioral model for the single-sideband mixer, we shall focus on the bands surrounding the fundamental RF tone. We are indeed concerned only with the harmonics and intermodulation products of the digital IF instead of the LO, since the LO harmonics can usually be easily removed using a lowpass filter as long as the IF frequency is a fraction of the LO frequency.

As we have clearly seen, even though the single-sideband mixer is supposed to generate only upper sidebands, we also need to account for the spurious lower sidebands generated. In view of the above discussion, we can then represent all the possible spurious bands around the fundamental by assuming that the IF signal is first amplified by a nonlinear IF amplifier that generates the harmonics of the IF signal before being up converted by an ideal I–Q modulator. Actually, two nonlinear IF amplifiers, U and L, are required, as is indicated on Figure 8.15, since we need to separately generate the lower and upper spurious sidebands at $\omega_{LO} \pm n\omega_{IF}$. Note that we assume that the bandwidth of the $I(t)$ and $Q(t)$ baseband signal is smaller than the IF frequency ω_{IF}. The I_{RF} and Q_{RF} signals for this model are therefore of the form

$$I_{RF}(t) = b_{IF}^{U}(t) + b_{IF}^{L}(t), \tag{8.7}$$

with

$$b_{IF}^{U/L}(t) = \sum_{i=0}^{n} I_i'^{U/L}(t)\cos(i\omega_{IF}t) - Q_i'^{U/L}(t)\sin(i\omega_{IF}t),$$

and

$$Q_{RF}(t) = \widehat{b_{IF}^{U}(t)} - \widehat{b_{IF}^{L}(t)}, \tag{8.8}$$

with

$$\widehat{b_{IF}^{U/L}(t)} = \sum_{i=0}^{n} I_i'^{U/L}(t)\sin(i\omega_{IF}t) + Q_i'^{U/L}(t)\cos(i\omega_{IF}t),$$

where $I_i'^{\mathrm{U/L}}$ and $Q_i'^{\mathrm{U/L}}$ are the functions introduced in Chapter 6 in Equation (8.9),

$$\begin{bmatrix} I_i'^{\mathrm{U/L}}(t) \\ Q_i'^{\mathrm{U/L}}(t) \end{bmatrix} = \begin{bmatrix} g_i^{\mathrm{U/L}}(E^2) & -h_i^{\mathrm{U/L}}(E^2) \\ h_i^{\mathrm{U/L}}(E^2) & g_i^{\mathrm{U/L}}(E^2) \end{bmatrix} \times \begin{bmatrix} I_i(t) \\ Q_i(t) \end{bmatrix}, \tag{8.9}$$

with $I_i(t) = E^i(t)\cos[i\theta(t)]$ and $Q_i(t) = E^i(t)\sin[i\theta(t)]$ the Chaillot functions defined using the envelope $E(t) = \sqrt{I^2 + Q^2}$ and phase $\theta = \angle(I + jQ)$. Note that superscripts U/L have been used to differentiate between the lower- and upper-sideband IF nonlinear amplifiers. The present model is a piecewise quasi-memoryless model but can be expanded to account for memory effects using the memory-polynomial technique introduced in Chapter 6.

Let us first use the behavioral model developed in the previous section for qualitatively explaining the strong spectral regrowth observed in Figure 8.14. In the experiment conducted in Figure 8.14, the wide bell-shaped spectral regrowth which takes place simultaneously at $\omega_{\mathrm{LO}} - 7\omega_{\mathrm{IF}}$, $\omega_{\mathrm{LO}} - 3\omega_{\mathrm{IF}}$, $\omega_{\mathrm{LO}} + \omega_{\mathrm{IF}}$ (in-band), and $\omega_{\mathrm{LO}} + 5\omega_{\mathrm{IF}}$ can be observed to arise rapidly above the noise floor once the amplitude of the input two-tone signal reaches a certain threshold. For lower input signal amplitudes, only the two and four center tones were observed, and the bell-shaped spectrum was not detected, since it was presumably buried below the noise floor. This suggests the presence of separate sources of nonlinearities competing in the transition from the weakly to the strongly nonlinear regime. Furthermore, this spectral regrowth appears to be similar to a phase-modulation process with a large phase-modulation index rather than an amplitude saturation/clipping, which would generate wider and flatter spectral regrowth. Note that the quadrature functions $g_i^{\mathrm{U/L}}[x^2(t)]$ and $h_i^{\mathrm{U/L}}[x^2(t)]$ introduced in Equation (8.9) can be expressed in terms of the amplitude function $f_i^{\mathrm{U/L}}[x^2(t)]$ and phase function $\phi_i^{\mathrm{U/L}}[x^2(t)]$ using

$$g_i^{\mathrm{U/L}}\left[x^2(t)\right] = f_i^{\mathrm{U/L}}\left[x^2(t)\right] \times \cos\left(\phi_i^{\mathrm{U/L}}[x^2(t)]\right),$$
$$h_i^{\mathrm{U/L}}\left[x^2(t)\right] = f_i^{\mathrm{U/L}}\left[x^2(t)\right] \times \sin\left(\phi_i^{\mathrm{U/L}}[x^2(t)]\right).$$

To account for the hard nonlinear regime observed, the nonlinear amplifiers U and L in the modulator model in Figure 8.15 can be approximated as the superposition of two nonlinear systems: the first features a dominant amplitude distortion, and the second features a dominant phase distortion. By implementing such a model in MATLAB using a ninth order Taylor series expansion for the phase $\phi_i[x^2(t)]$, we can check the capability of the model to operate both in the weakly and in the strongly nonlinear regime. The simulation results for the strongly nonlinear regime are shown in Figure 8.16. At higher input signal power levels, the phase-distortion products (stem line with crosses) resulting from the phase modulation $\phi_i[x^2(t)]$ become comparable to the amplitude-distortion products (stem lines with circles) and in some cases (such as $\omega_{\mathrm{LO}} - 7\omega_{\mathrm{IF}}$ and $\omega_{\mathrm{LO}} + 5\omega_{\mathrm{IF}}$) overcome it.

Under normal operation of the *I–Q* modulator, the hard nonlinear regime is avoided, and in the next section we shall limit our analysis to the weakly nonlinear regime.

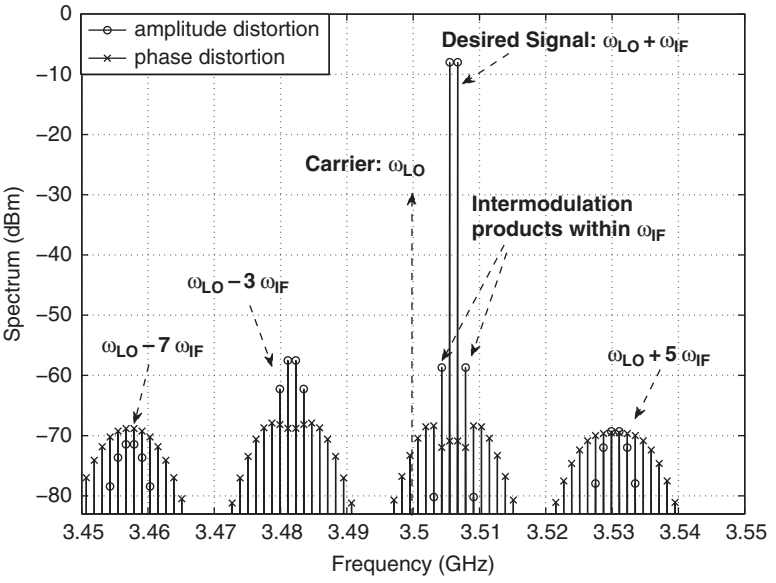


Figure 8.16 Results for a mixer model with amplitude and phase distortion. (MATLAB simulation performed by Xi Yang at Ohio State University.)

8.3.2 Poly-harmonic predistortion linearization test results

Having developed a behavioral model to predict the nonlinear response of the modulator, we will now focus on linearizing it using predistortion. The conceptual problem of inverting a transfer function will be discussed in the next chapter. For our present purposes we shall proceed empirically and postulate that the modulator is operated in its weakly nonlinear regime so that we can assume that the weak predistortion corrections signals are themselves principally quasi-linearly up converted at the output of the mixer.

For mild nonlinearities the linearization functions $g_i^{\mathrm{U/L}}[x^2(t)]$ and $h_i^{\mathrm{U/L}}[x^2(t)]$ can be well implemented using a power series. The Taylor series parameters can be systematically extracted for gradually increasing input power levels by tuning the amplitude and phase of each coefficient. Owing to the frequency selectivity of this approach, only one complex coefficient needs to be tuned at a time. Alternatively, an adaptive extraction scheme can be implemented. At any rate, the goal is to implement a linearization scheme that works for a wide range of power levels in the weakly nonlinear regime of the single-sideband mixer.

Figure 8.17(a) and (b) shows the output of the single-sideband mixer (four paths) for a two-tone excitation before and after poly-harmonic predistortion linearization. The input baseband signal is up converted to 3.5 GHz. In Figure 8.17(b) the LO leakage was removed using a DC offset at the I and Q inputs. The image band was suppressed using balancing, and the in-band intermodulation products and the IF harmonics were removed using the poly-harmonic predistortion algorithm.

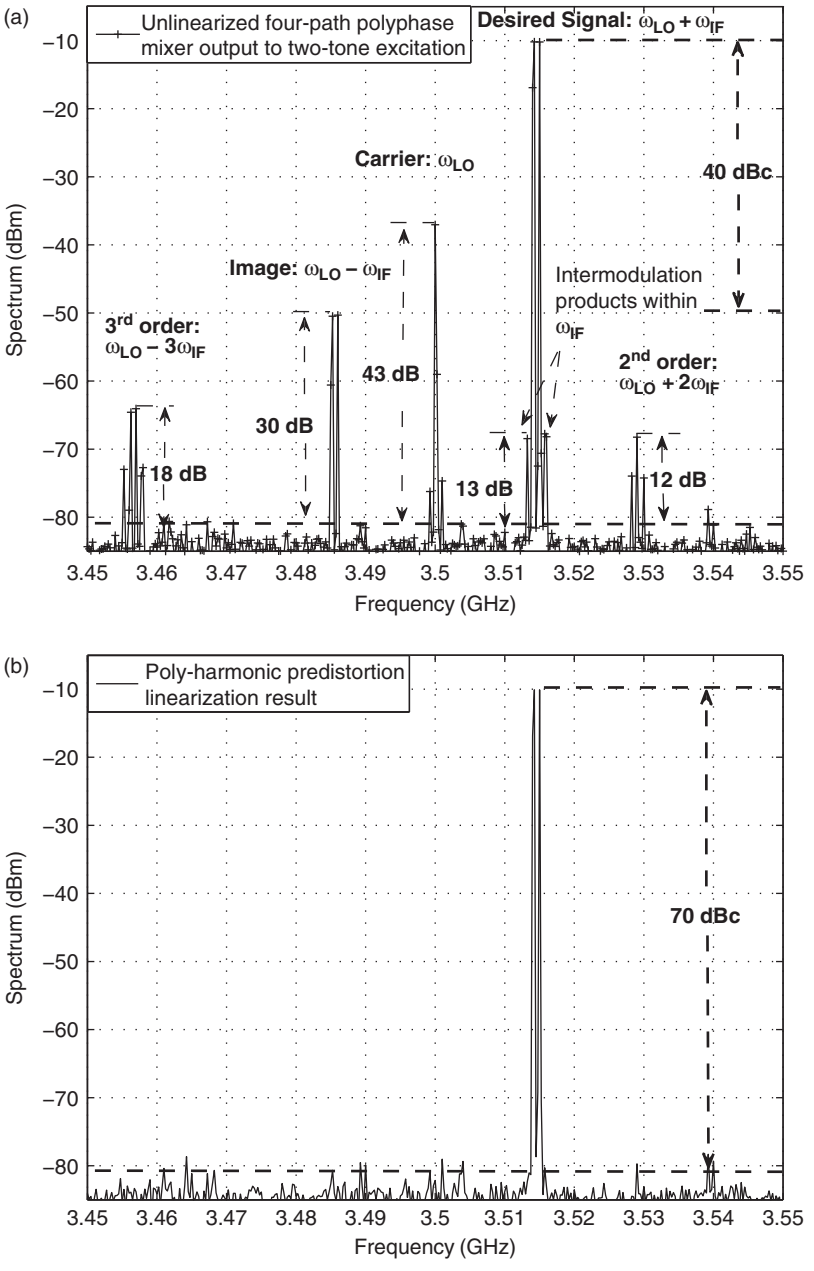


Figure 8.17 Poly-harmonic predistortion linearization results of a four-path polyphase mixer in response to a two-tone input signal measured with a 30-kHz resolution bandwidth. (From [12] with permission, ©2010 IEEE.)

When the input signal power level is increased up to the limit of the strongly nonlinear regime, the distortions remain canceled out without any further tuning of the predistorted coefficients. In this particular experimental investigation the poly-harmonic predistortion could handle an input signal power range up to −3 dBm for 2-tone and

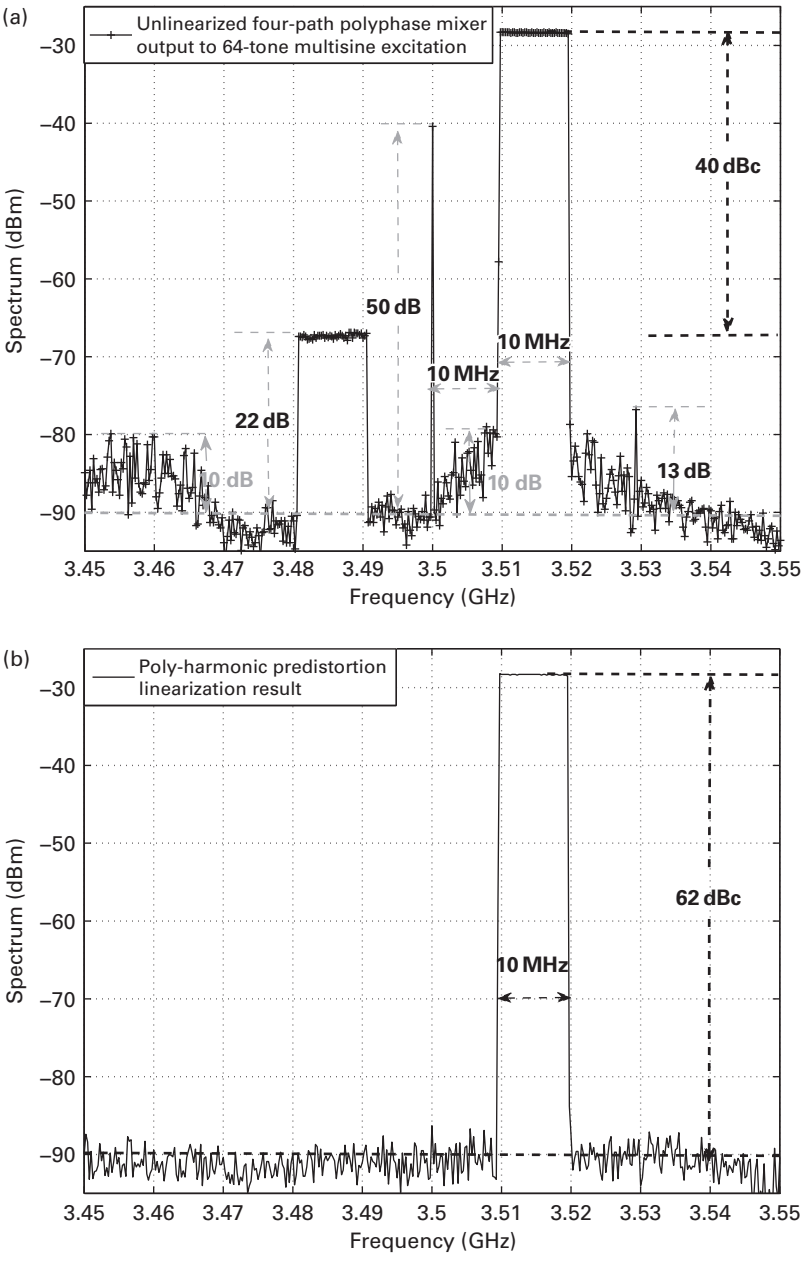


Figure 8.18 Poly-harmonic predistortion linearization results of a four-path polyphase mixer in response to a 64-tone input signal measured with a 10-kHz resolution bandwidth. (From [12] with permission, ©2010 IEEE.)

−6.2 dBm for 64-tone multisine excitations, respectively, for the device under test (TRF3703).

The results for a 64-tone multisine input signal of bandwidth 10 MHz before and after poly-harmonic predistortion linearization are shown in Figures 8.18(a) and (b). In

Table 8.1. Summary of results obtained

Input	Third harmonic	Image	LO	In-band	Second harmonic	ACLR
Two-tone	18 dB	30 dB	43 dB	13 dB	12 dB	−70 dBc
Multisine	10 dB	22 dB	50 dB	10 dB	13 dB	−62 dBc
OFDM	8 dB	19 dB	50 dB	–	6 dB	−60 dBc

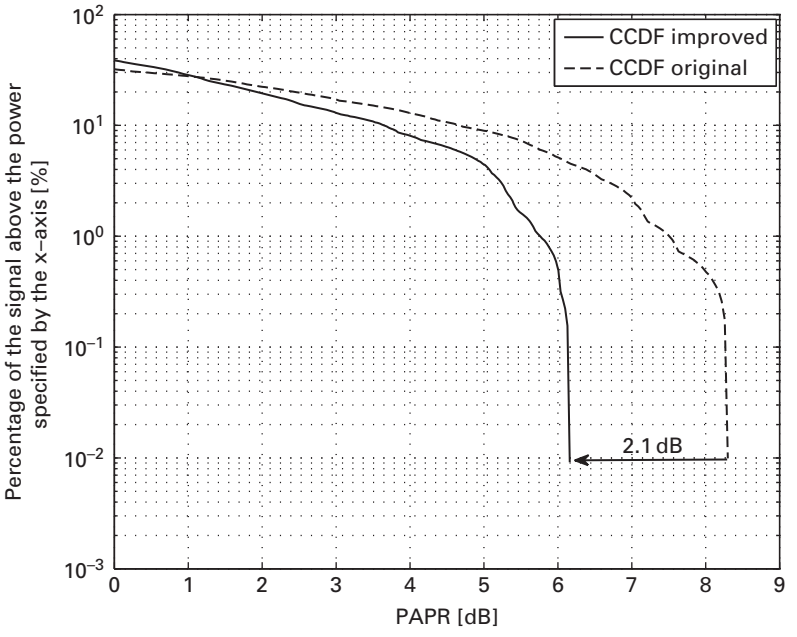


Figure 8.19 A CCDF plot of the 64-tone baseband multisine signal. (From [12] with permission, ©2010 IEEE.)

this linearization process, the predistortion coefficients used are the same as those used for the two-tone linearization in Figure 8.18(b). This indicates that the poly-harmonic predistortion linearization works with a multi-carrier modulation input signal with high envelope fluctuations. Similar results are also obtained using an OFDM signal [11].

It can be seen that the fewer the input tones, the higher the adjacent-channel leakage ratio (ACLR) the system can achieve. This is to be expected since the more input tones, the higher the peak-to-average power ratio (PAPR). The multisine CCDF[3] was optimized (see Figure 8.19) so as to approximate the CCDF of the OFDM signal. The signal selected features a maximum PAPR of 6.1 dB with a probability of 0.01%.

Table 8.1 summarizes the results obtained using the various types of signals used. Listed are the reductions achieved for the third-order harmonic, image band, LO

[3] See Chapter 1 for a definition.

leakage, in-band ACLR, and second harmonic as well as the overall spurious rejection. The overall rejections vary from −70 dBc to −60 dBc depending on the signal type considered.

In summary, the poly-harmonic predistortion linearization technique can be used to optimize the output spectrum of a polyphase multi-path circuit operated in the weakly nonlinear regime. The polyphase multi-path technique combined with poly-harmonic predistortion linearization offers an attractive approach for the development of *filterless* software-defined radios.

References

[1] Jian Lin and E. Tsui, "Joint adaptive transmitter/receiver *IQ* imbalance correction for OFDM systems," in *IEEE International Symposium on PIMRC*, Vol. 2, pp. 1511–1516, 2004.

[2] Kuang-Hao Lin, Hsin-Lei Lin, Shih-Ming Wang, and Chang R. C.,"Implementation of digital *IQ* imbalance compensation in OFDM WLAN receivers," in *IEEE International Symposium on Circuits and Systems*, 2006.

[3] P. Roblin, Sukkeun Myoung, D. Chaillot, Young Gi Kim, A. Fathimulla, J. Strahler, and S. Bibyk, "Frequency selective predistortion linearization of RF power amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 56, No. 1, pp. 65–76, Jan. 2008.

[4] Behzad Razavi, *RF Microelectronics*, Prentice Hall, 1998.

[5] L. Moult and J. E. Chen, "The K-model: RFIC modeling for communication systems simulation," in *IEE Colloquium on Analog Signal Processing*, pp. 11/1–11/8, 1998.

[6] W. V. Moer and Y. Rolain, "Determining the reciprocity of mixers through 3-port large signal network analyzer measurement", in *62nd ARFTG Conference Digest*, pp. 165–170, 2003.

[7] S. K. Myoung, X. Cui, P. Roblin, D. Chaillot, F. Verbeyst, M. Vanden Bossche, S. J. Doo, and W. Dai, "Large signal network analyzer with trigger for baseband modulation linearization," in *64th ARFTG Conference Digest*, Orlando, FL, 2004.

[8] S. Mutha, P. Roblin, D. Chaillot, X. Yang, K. Jiwoo, J. Strahler, R. Rojas-Tehra, and J. Volakis, "Technique for joint balancing of *IQ* modulator–demodulator chains in wireless transmitters," in *2009 Microwave Symposium Digest*, pp. 221–224, 2009.

[9] E. Mensink, E. A. M. Klumperink, and B. Nauta, "Distortion cancellation by polyphase multipath circuits," *IEEE Transactions on Circuits and Systems I*, Vol. 52, No. 9, pp. 1785–1794, Sept. 2005.

[10] R. Shrestha, E. A. M. Klumperink, E. Mensink, G. J. M. Wienk, and B. Nauta, "A polyphase multipath technique for software-defined radio transmitters," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 12, pp. 2681–2692, Dec. 2006.

[11] D. K. Weaver, "A third method of generation and detection of single sideband signals," *Proceedings of the IRE*, Vol. 44, No. 12, pp. 1703–1705, Dec. 1956.

[12] X. Yang, D. Chaillot, P. Roblin, W.-R. Liou, J. Lee, H.D. Park, J. Strahler, and M. Ismail, "Poly-harmonic modeling and predistortion linearization for software defined radio upconverters," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 58, No. 8, pp. 2125–2133, Aug. 2010.

9 Linearization of RF power amplifiers with memory[1]

One of the most challenging issues in designing RF power amplifiers is the linearity requirement. The spurious emissions from nonlinear RF power amplifiers are spread out over neighboring channels. As more complex wideband modulation techniques such as wideband CDMA and OFDM are used and also combined in multi-carrier and even multiband transmitters, higher peak-to-average power ratios (PAPRs) (e.g. 4.5 dB for a handset and 12 dB for a basestation in WCDMA) result, imposing stronger linearity requirements on RF PAs. Memory effects also become more significant in high-efficiency PAs operating with wideband signals and need to be taken into account for their linearization. As we shall see, memory effects can be classified into two main types: slow memory effects and fast memory effects [1] [2]. Slow memory effects, which usually encompass temperature effects [3], traps, and power supplies' response, can usually be dealt with using adaptive linearization techniques [4] [5]. On the other hand, fast memory effects, which typically originate from the intrinsic transistor as well as matching and bias networks, are usually observed above 1 MHz [2] and require more advanced instantaneous linearization techniques.

This chapter will focus on predistortion linearization for a few canonic cases. First we will demonstrate how predistortion linearization is affected by electrical and self-heating memory effects. Next we will investigate the linearization of quasi-memory-less PA exhibiting AM/AM and AM/PM distortion. The linearization of PA with memory will then be studied for the case of PAs that are well modeled using memory polynomials [6] [7]. We will then conclude this chapter with the presentation of a frequency-selective two-band predistorter [8]. These examples are intended to demonstrate the close relationship between PA behavioral modeling and predistortion linearization.

9.1 Predistortion linearization and the impact of memory effects

In Chapter 4 we saw that a multi-*R*–*C*-time-constant electrical network could be used to model the transient thermal response of an RF transistor. Combining such a thermal

[1] Research collaboration with Suk Keun Myoung, Wenhua Dai, Dominique Chaillot, and Xi Yang is gratefully acknowledged.

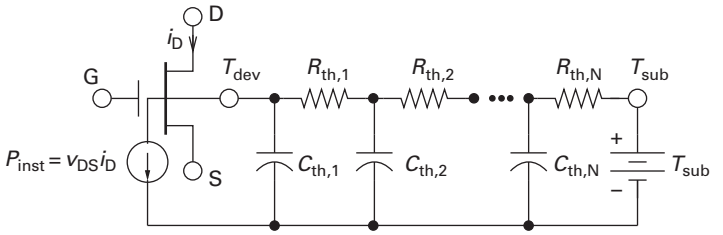


Figure 9.1 Electrothermal FET model for thermal transient simulations.

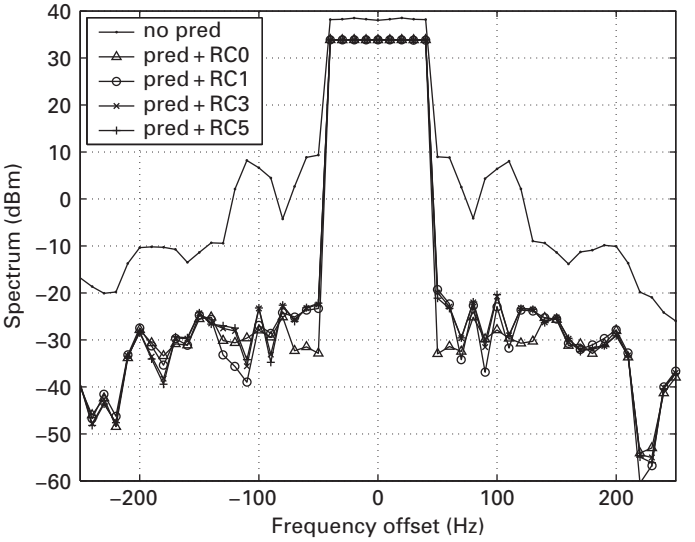


Figure 9.2 Output spectra of a 60-W class-AB amplifier for various thermal models, tone spacing at 10 Hz. (ADS simulation by Wenhua Dai at Ohio State University.)

network with a temperature-dependent device (see Figure 9.1) provides us with a realistic electrothermal model for predicting the transient thermal response of a power transistor excited by a time-varying electrical excitation. Using such an electrothermal model, we shall now investigate the impact of both thermal and electrical memory effects on a linearized amplifier.

For this purpose we shall use a simulation experiment in which an LDMOSFET transistor is implemented with four different thermal networks in a circuit simulator. The four non-distributed models are referred to as RC0 ($R_{\text{th}} = 0$), RC1 (one-stage RC), RC3 (three-stage RC), and RC5 (five-stage RC) depending on the number of R–C elements involved. RC0 is an isothermal model, since the device temperature remains constant. All of the other models have the same total steady-state $R_{\text{th}} = \sum_i R_{\text{th},i}$. A 60-W class-AB power amplifier was then designed using the RC1 transistor model.

To approximate a communication signal with a wide time-varying envelope, a nine-tone multi-sine excitation with a PAPR of 6.8 dB is used. The spectrum of the transistor output obtained for this multisine is shown in Figure 9.2.

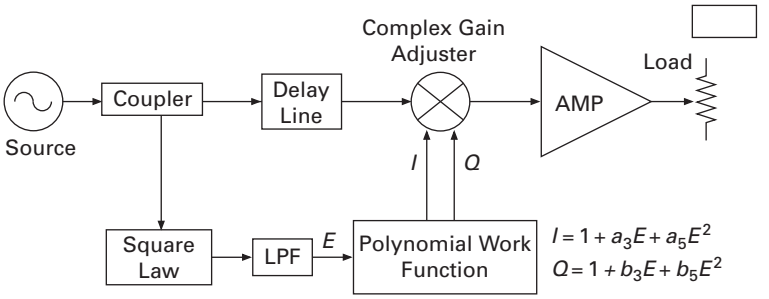


Figure 9.3 A schematic diagram of an amplifier with RF predistortion.

The goal of our study is to evaluate the dependence of the ACPR of a linearized amplifier upon the bandwidth of the multisine for the various thermal network models RC0 to RC5. The ACPR in these multitone simulations is defined as the ratio of the *total adjacent-band power* to the *total in-band power*. An RF predistorter circuit (Figure 9.3) was thus implemented to linearize the amplifier designed. The predistorter improves the ACPR, but at the same time also makes the amplifier more sensitive to memory effects. For the amplifier under study, it increases the sensitivity (derivative) of the ACPR to the device temperature by a factor of 10. Note that the bias tees used also have a strong impact on the performance of the predistortion linearization. Indeed, the modulated RF signals are down converted by the device's electrical nonlinearities. It results that a baseband signal (square of the RF envelope; loosely referred to as IF) can arise at the drain terminal. The impedance termination presented by the bias tee at IF will then consequently greatly impact the nonlinear response (ACPR) of the amplifier for varying RF bandwidths. To help with the separation of the electrical memory effects from the thermal memory effects, a bias tee that shorts the down-converted IF electrical signals was implemented in the simulations, suppressing their impact on the ACPR. However, the IF fluctuations of the temperature induced by the instantaneous power dissipated (Joule effect, $i_D v_{DS}$) will continue to modulate the device characteristics and impact its ACPR.

To investigate the thermal effects, the predistorter coefficients are first optimized for a tone-spacing of 0.1 MHz (corresponding to a signal bandwidth of 0.9 MHz) using the RC1 model. The input power used is set so that the average output power is 6 dB backed off from P1dB. The predistorter improves the ACPR by about 30 dB.

Next the tone-spacing is swept from 1 Hz for 1 MHz and the lower-sideband (LSB) and upper-sideband (USB) ACPR are measured. Four different thermal models (RC0, RC1, RC3, and RC5) are then used, and the ACPR versus bandwidth is measured and compared for each thermal model.

The LSB ACPR results are displayed in Figures 9.4(a) and (b). The results for models RC3 and RC5 are similar, indicating that the transistor response is accurately represented with a network relying on three thermal *R–C* time constants.

Among the thermal models compared in Figure 9.4, the RC0 model is the only one which cannot exhibit thermal memory effects, since it has a thermal resistance

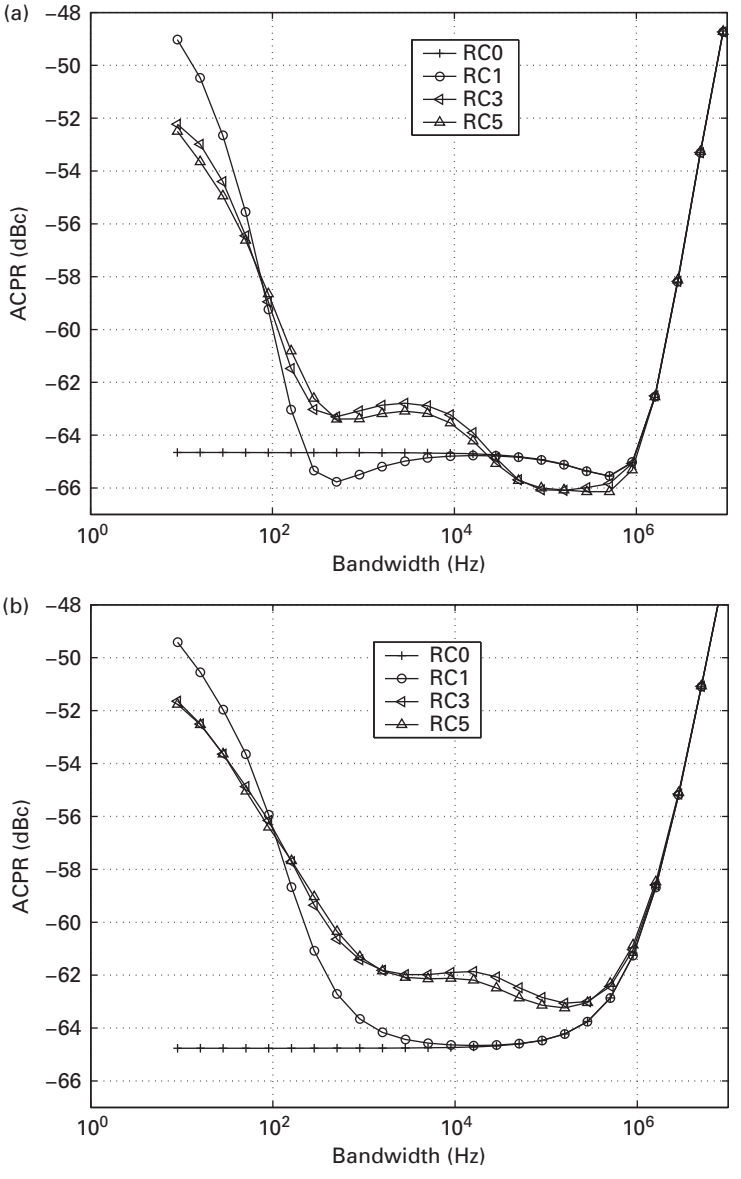


Figure 9.4 (a) Lower-sideband and (b) upper-sideband ACPR versus bandwidth for a 60-W amplifier under nine-tone excitation. (ADS simulation by Wenhua Dai at Ohio State University.)

of zero, and there is no temperature fluctuation with the signal envelope (the constant device temperature in RC0 is simply set to the average temperature of RC1 at 0.1 MHz tone spacing). It results that the memory effects in RC0 are purely of electrical origin. Any deviations arising in other thermal models are therefore of thermal origin. It is observed in Figure 9.4 that, above 1 MHz bandwidth, all thermal models exhibit the same magnitude of ACPR degradation. This indicates that the *electrical* memory

effects completely dominate the *thermal* memory effects for signals with bandwidth above 1 MHz. However, below 1 MHz bandwidth the ACPR of RC0 does not change much, whereas other thermal models exhibit a greater degradation of the ACPR. This indicates that thermal memory effects dominate electrical memory effects for signal bandwidth below 1 MHz.

Below 1 MHz bandwidth, when thermal memory effects dominate over electrical memory effects, the ACPR degradation remains moderate down to 100 Hz bandwidth. But for bandwidth below 100 Hz the ACPR degradation rapidly rises again and reaches 17 dB at 10 Hz bandwidth. Obviously, in practice the signal bandwidth is not normally that small. Instead, these slow thermal effects will be excited by slow variations in the average power dissipated by the amplifier.

Above 1 MHz bandwidth, when electrical memory effects dominate over thermal memory effects, the ACPR degrades by up to 17 dB as the bandwidth approaches 10 MHz. Clearly, due to memory effects, the linearization performance of a memoryless predistorter will degrade with increasing bandwidth.

9.2 Predistortion for quasi-memoryless amplifiers

As we have seen, memory effects in power amplifiers cannot be neglected in predistortion linearization since they degrade the linearization performance. To address this issue, we first consider the case of an amplifier for which only quasi-memoryless (QML) distortions are accounted for.

Let us assume that the RF amplifier is excited by a modulated RF signal x_{in}:

$$x_{\text{in}} = I_{\text{in}}(t)\cos(\omega t) - Q_{\text{in}}(t)\sin(\omega t),$$

with I_{in} and Q_{in} the baseband modulation signals. The output $b_{\text{out}}^{\text{QML}}$ of the QML power amplifier excited by x_{in} can then be expressed as follows:

$$b_{\text{out}}^{\text{QML}}(t) = I_{\text{A,out}}(t)\cos(\omega_0 t) - Q_{\text{A,out}}(t)\sin(\omega_0 t),$$

where $I_{\text{A,out}}$ and $Q_{\text{A,out}}$ are given by

$$\begin{bmatrix} I_{\text{A,out}}(t) \\ Q_{\text{A,out}}(t) \end{bmatrix} = \begin{bmatrix} G(E^2) & -H(E^2) \\ H(E^2) & G(E^2) \end{bmatrix} \times \begin{bmatrix} I_{\text{in}}(t) \\ Q_{\text{in}}(t) \end{bmatrix} = \mathbf{K}_{\text{A}} \times \begin{bmatrix} I_{\text{in}}(t) \\ Q_{\text{in}}(t) \end{bmatrix},$$

with $G(E^2)$ and $H(E^2)$ two functions of the envelope squared, $E^2 = I_{\text{in}}^2 + Q_{\text{in}}^2$, and $\mathbf{K}_{\text{A}}(E^2)$ the nonlinear K-matrix for the PA. The functions $G(E^2)$ and $H(E^2)$ can themselves be expressed as

$$G(E^2) = F(E^2)\cos\left[\phi_{\text{A}}(E^2)\right],$$
$$H(E^2) = F(E^2)\sin\left[\phi_{\text{A}}(E^2)\right],$$

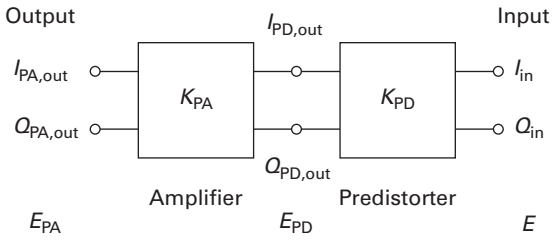


Figure 9.5 K-matrix cascade for predistortion linearization.

where the amplitude $F(E^2)$ and phase $\phi_A(E^2)$ functions describe the AM/AM and AM/PM distortion. The gain $F(E^2)$ and phase $\phi_A(E^2)$ functions satisfy the property

$$F(E^2) = \left[G^2(E^2) + H^2(E^2)\right]^{1/2} = \frac{\left[I_{A,out}^2 + Q_{A,out}^2\right]^{1/2}}{\left[I_{in}^2 + Q_{in}^2\right]^{1/2}} = \frac{E_{A,out}}{E},$$
$$\phi_A(E^2) = \angle\left[G(E^2) + jH(E^2)\right] = \angle\left(I_{A,out} + jQ_{A,out}\right) - \angle\left(I_{in} + jQ_{in}\right).$$

In the limit of small input power, the PA is linear, and the nonlinear gain reduces asymptotically to the linear PA gain $G_A^{lin,0}$:

$$F(E^2) \simeq G_A^{lin,0}.$$

We now wish to linearize the PA by predistorting its input signal. This is represented symbolically in Figure 9.5 by cascading the predistortion block (PD) and the power amplifier (PA).

Since the PA is QML, we can infer that the required predistortion function will be QML. The output b_{out}^P of the QML predistorter excited by x_{in} can then be expressed as follows:

$$b_{out}^{QML}(t) = I_{P,out}(t)\cos(\omega_0 t) - Q_{P,out}(t)\sin(\omega_0 t),$$

where $I_{P,out}$ and $Q_{P,out}$ are given by

$$\begin{bmatrix} I_{P,out}(t) \\ Q_{P,out}(t) \end{bmatrix} = \begin{bmatrix} g(E^2) & -h(E^2) \\ h(E^2) & g(E^2) \end{bmatrix} \times \begin{bmatrix} I_{in}(t) \\ Q_{in}(t) \end{bmatrix} = \mathbf{K}_P(E^2) \times \begin{bmatrix} I_{in}(t) \\ Q_{in}(t) \end{bmatrix},$$

with $\mathbf{K}_P(E^2)$ the nonlinear K-matrix for the predistorter.

The role of the predistorter is to compensate for the gain compression of the PA so that the gain of the PA and PD cascaded remains constant. This implies a reduction of the PA gain at low input power by a multiplicative factor α smaller than unity, as is illustrated in Figure 9.6. Multiplying the K-matrices of the PA and PD system thus yields the K-matrix for the linearized PA with gain $\alpha G_A^{lin,0}$:

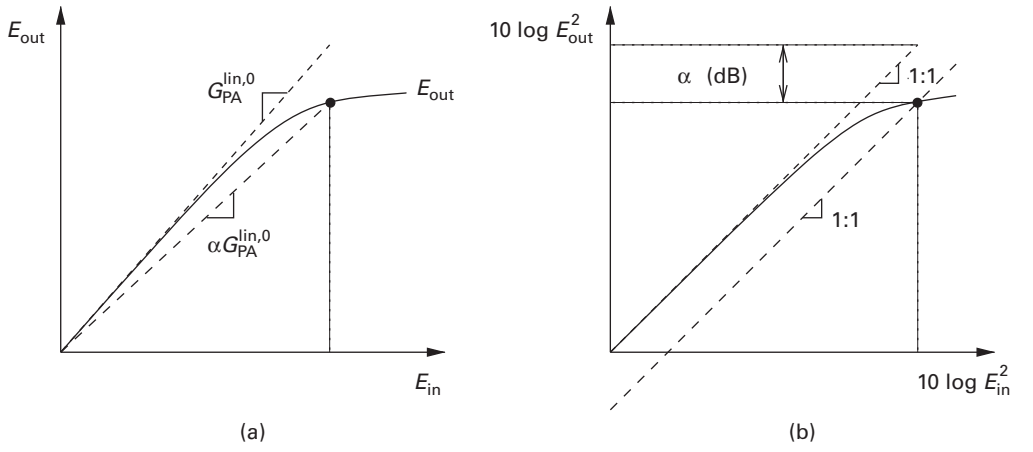


Figure 9.6 Gain reduction in predistortion linearization on linear (a) and logarithmic (b) scales.

$$\mathbf{K}_{\mathrm{A}}(E_{\mathrm{P}}^2) \times \mathbf{K}_{\mathrm{P}}(E^2) = \begin{bmatrix} G(E_{\mathrm{P}}^2) & -H(E_{\mathrm{P}}^2) \\ H(E_{\mathrm{P}}^2) & G(E_{\mathrm{P}}^2) \end{bmatrix} \times \begin{bmatrix} g(E^2) & -h(E^2) \\ h(E^2) & g(E^2) \end{bmatrix},$$
$$= \alpha G_{\mathrm{A}}^{\mathrm{lin},0} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{9.1}$$

where we have defined $E_{\mathrm{P}} = \sqrt{I_{\mathrm{P,out}}^2 + Q_{\mathrm{P,out}}^2}$ as the envelope at the predistorter output. Equation (9.1) leads to the following system of equations:

$$\begin{aligned} G\left(E_{\mathrm{P}}^2\right) g(E^2) - H\left(E_{\mathrm{P}}^2\right) h(E^2) &= \alpha G_{\mathrm{A}}^{\mathrm{lin},0}, \\ G\left(E_{\mathrm{P}}^2\right) h(E^2) + H\left(E_{\mathrm{P}}^2\right) g(E^2) &= 0. \end{aligned} \tag{9.2}$$

This system admits the following solution:

$$\begin{aligned} g(E^2) &= \alpha G_{\mathrm{A}}^{\mathrm{lin},0} \frac{G\left(E_{\mathrm{P}}^2\right)}{F^2\left(E_{\mathrm{P}}^2\right)}, \\ h(E^2) &= -\alpha G_{\mathrm{A}}^{\mathrm{lin},0} \frac{H\left(E_{\mathrm{P}}^2\right)}{F^2\left(E_{\mathrm{P}}^2\right)}, \end{aligned} \tag{9.3}$$

with

$$E_{\mathrm{P}} \cdot F\left(E_{\mathrm{P}}^2\right) = E_{\mathrm{A}} = \alpha G_{\mathrm{A}}^{\mathrm{lin},0} E.$$

Note that the in-band envelopes $E^2 = I_{\mathrm{in}}^2 + Q_{\mathrm{in}}^2$, $E_{\mathrm{P}}^2 = I_{\mathrm{P,out}}^2 + I_{\mathrm{P,out}}^2$, and $E_{\mathrm{A}}^2 = I_{\mathrm{A,out}}^2 + I_{\mathrm{A,out}}^2$ are those measured at the input and output of the predistortion stage and the output of the amplifier, respectively. For the solution to hold, Equation (9.3) needs to be solved for E_{P} for all input E. Since Equation (9.3) gives E in terms of E_{P}, we simply need to invert it. A unique solution will therefore exist when the nonlinear function $E_{\mathrm{P}} \cdot F(E_{\mathrm{P}}^2)$ is a monotonic function of E_{P}.

9.3 Linearization for PAs modeled with memory polynomials

We have seen in Chapter 6 that a popular behavioral model accounting for memory effects relies on the memory-polynomial approximation in which only the diagonal memory effects are accounted for in the Volterra series expansion [6] [7]. In discrete time, this yields the following K-matrix system:

$$\begin{bmatrix} I_{\mathrm{A}}(n) \\ Q_{\mathrm{A}}(n) \end{bmatrix} = \sum_{i\geq 0}^{N_i} \left\{ \begin{bmatrix} G_i \left[E_{\mathrm{P}}^2(n-i)\right] & -H_i \left[E_{\mathrm{P}}^2(n-i)\right] \\ H_i \left[E_{\mathrm{P}}^2(n-i)\right] & G_i \left[E_{\mathrm{P}}^2(n-i)\right] \end{bmatrix} \times \begin{bmatrix} I_{\mathrm{P}}(n-i) \\ Q_{\mathrm{P}}(n-i) \end{bmatrix} \right\},$$

where n is the time index, i the memory delay index, and $N_i + 1$ the total number of memory taps required to model the PA. In a non-matrix form this can be rewritten as

$$I_{\mathrm{A}}(n) = \sum_{i\geq 0}^{N_i} \left\{ G_i \left[E_{\mathrm{P}}^2(n-i)\right] I_{\mathrm{P}}(n-i) - H_i \left[E_{\mathrm{P}}^2(n-i)\right] Q_{\mathrm{P}}(n-i) \right\},$$

$$Q_{\mathrm{A}}(n) = \sum_{i\geq 0}^{N_i} \left\{ H_i \left[E_{\mathrm{P}}^2(n-i)\right] I_{\mathrm{P}}(n-i) + G_i \left[E_{\mathrm{P}}^2(n-i)\right] Q_{\mathrm{P}}(n-i) \right\}.$$

We wish now to linearize the PA by predistorting the input signal. To do so we shall assume that the predistorted signal can be quite generally expanded in terms of the various delayed input signals $I(n-j)$ and $Q(n-j)$ as

$$I_{\mathrm{P}}(n) = \sum_{j\geq 0}^{N_j} \left\{ g_j(n) I_{\mathrm{P}}(n-j) - h_j(n) Q_{\mathrm{P}}(n-j) \right\}, \tag{9.4}$$

$$Q_{\mathrm{P}}(n) = \sum_{i\geq 0}^{N_j} \left\{ h_j(n) I_{\mathrm{P}}(n-j) + g_j(n) Q_{\mathrm{P}}(n-j) \right\}, \tag{9.5}$$

with $N_j + 1$ the total number of memory taps to be used by the predistorter. Note that the PD functions $g_j(n)$ and $h_j(n)$ are not dependent on the input envelope $E(n)$, but rather are time-varying functions to be determined at each time n as a function of the present ($j = 0$) and past ($j > 0$) values of $I(n-j)$ and $Q(n-j)$.

On substituting the predistortion output in to the PA model we obtain the following equations for $I_{\mathrm{A}}(n)$:

$$\begin{aligned} I_{\mathrm{A}}(n) = & \sum_{j\geq 0}^{N_j} \sum_{i\geq 0}^{N_i} \left\{ G_i \left[E_{\mathrm{P}}(n-i)\right] g_j(n-i) - H_i \left[E_{\mathrm{P}}(n-i)\right] h_j(n-i) \right\} I[n-i-j] \\ & - \sum_{j\geq 0}^{N_j} \sum_{i\geq 0}^{N_i} \left\{ G_i [E_{\mathrm{P}}(n-i)] h_j(n-i) + H_i [E_{\mathrm{P}}(n-i)] g_j(n-i) \right\} Q[n-i-j]. \end{aligned}$$

A similar redundant equation is obtained for $Q_{\mathrm{A}}(n)$. On setting to zero all the coefficients weighting $I[n-i-j]$ and $Q[n-i-j]$ except for the case $i + j = i = j = 0$, which is for the linearized PA gain, we obtain

$$\sum_{i\geq\max(0,k-N_j)}^{\min(k,N_i)} G_i\left[E_{\mathrm{P}}(n-i)\right]g_{k-i}(n-i) - H_i\left[E_{\mathrm{P}}(n-i)\right]h_{k-i}(n-i) = \alpha G_{\mathrm{A}}^{\mathrm{lin},0}\delta_{k,0} \quad \forall\ 0\leq k\leq N_j+N_i,$$

$$\sum_{i\geq\max(0,k-N_j)}^{\min(k,N_i)} G_i\left[E_{\mathrm{P}}(n-i)\right]h_{k-i}(n-i) + H_i\left[E_{\mathrm{P}}(n-i)\right]g_{k-i}(n-i) = 0 \quad \forall\ 0\leq k\leq N_j+N_i.$$

The linearization of a PA model with N_i+1 taps leads to a system of $2(N_i+N_j+1)$ equations for the $I(n-i-j)$ and $Q(n-i-j)$ input signals, with $2(N_j+1)$ unknown predistorter coefficients $g_j(n)$ and $h_j(n)$ to determine at time n. Clearly, this is an overdetermined system. Only the equations from $k=0$ up to $k=N_i$ can be solved and the equations from $k=N_i+1$ to $k=N_i+N_j$ introduce a residual error. However, under normal conditions this residual error decreases with the number of taps N_j+1 in the predistorter. An exact solution can then be approached if we let N_j+1 (the number of taps of the predistorter) go to infinity. In practice it is found, however, that satisfactory convergence occurs with a small number of taps in the predistorter.

Let us consider the case of a PA in which a couple of taps (one discrete delay) is sufficient to account for the memory effects in the PA. Using the variables

$$F_0^2 = \left[E_{\mathrm{P}}^2(n)\right]G_0^2\left[E_{\mathrm{P}}^2(n)\right] + H_0^2\left[E_{\mathrm{P}}^2(n)\right],$$

$$A(n) = -\frac{G_0\left[E_{\mathrm{P}}^2(n)\right]G_1\left[E_{\mathrm{P}}^2(n-1)\right] + H_0\left[E_{\mathrm{P}}^2(n)\right]H_1\left[E_{\mathrm{P}}^2(n-1)\right]}{F_0^2\left[E_{\mathrm{P}}^2\right]},$$

$$B(n) = \frac{H_0\left[E_{\mathrm{P}}^2(n)\right]G_1\left[E_{\mathrm{P}}^2(n-1)\right] - G_0\left[E_{\mathrm{P}}^2(n)\right]H_1\left[E_{\mathrm{P}}^2(n-1)\right]}{F_0^2\left[E_{\mathrm{P}}^2\right]},$$

the required predisorter functions $g_j(n)$ and $h_j(n)$ are then given by

$$g_0(n) = \alpha G_{\mathrm{A}}^{\mathrm{lin},0}\frac{G_0\left[E_{\mathrm{P}}^2(n)\right]}{F_0^2\left[E_{\mathrm{P}}^2(n)\right]}, \tag{9.6}$$

$$h_0(n) = -\alpha G_{\mathrm{A}}^{\mathrm{lin},0}\frac{H_0\left[E_{\mathrm{P}}^2(n)\right]}{F_0^2\left[E_{\mathrm{P}}^2(n)\right]}, \tag{9.7}$$

$$g_k(n) = A(n)g_{k-1}(n-1) - B(n)h_{k-1}(n-1) \qquad \forall\ 0\leq k\leq N_j, \tag{9.8}$$

$$h_k(n) = B(n)g_{k-1}(n-1) + A(n)h_{k-1}(n-1) \qquad \forall\ 0\leq k\leq N_j. \tag{9.9}$$

The first two equations are the same as in the QML case. The remaining equations are associated with memory predistorter corrections. Note that the functions $g_j(n)$ and $h_j(n)$ are simultaneously dependent, via the parameters $A(n)$ and $B(n)$, on both the present ($k=0$) and the past ($k>0$) values of the envelope $E_{\mathrm{P}}(n-k)$. It results that they cannot be expressed in terms of a single $E_{\mathrm{P}}(n-i)$, and the predistorter cannot be recast in terms of memory polynomials like the PA.

A key feature of the solution expressed by Equations (9.6)–(9.9) is that the envelope $E_{\mathrm{P}}(n)$ at the output of the predistorter is unknown at time n and must be self-consistently calculated at each new time step:

$$E_{\mathrm{P}}(n) = \left[I_{\mathrm{P}}^2(n) + Q_{\mathrm{P}}^2(n)\right]^{1/2} f_n[E_{\mathrm{P}}(n)]. \tag{9.10}$$

Indeed, $I_{\mathrm{P}}(n)$ and $Q_{\mathrm{P}}(n)$ are calculated using Equations (9.4) and (9.5) from known past values of $I_{\mathrm{P}}(n-i)$ and $Q_{\mathrm{P}}(n-i)$ $(i>1)$ and yet-to-be-determined current values of $g_k(n)$ and $h_k(n)$. But the values of $g_k(n)$ and $h_k(n)$, which are functions of past values $g_k(n-i)$ and $h_k(n-i)$ $(i>1)$ in Equations (9.6)–(9.9), are also functions of the yet-to-be-determined envelope $E_{\mathrm{P}}(n)$ via the $A(n)$ and $B(n)$ variables. Thus Equation (9.10) defined a time-varying transcendental equation $E_{\mathrm{P}} = f_n[E_{\mathrm{P}}(n)]$ in terms of $E_{\mathrm{P}}(n)$ that needs to be solved at each time step n.

The generalization of the above results to an arbitrarily large number of taps N_i+1 and N_j+1 is more readily conducted using a complex-number representation:

$$\begin{aligned} X_A(n) &= I_{\mathrm{A,out}}(n) + jQ_{\mathrm{A,out}}(n), \qquad X_{\mathrm{P}}(n) = I_{\mathrm{P,out}}(n) + jQ_{\mathrm{P,out}}(n), \\ X(n) &= I(n) + jQ(n), \\ \mathcal{K}_i\left[E_{\mathrm{P}}^2\right] &= G_i\left[E_{\mathrm{P}}^2\right] + jH_i\left[E_{\mathrm{P}}^2\right], \qquad \mathcal{L}_j(n) = g_j(n) + jh_j(n). \end{aligned}$$

The PA output is then rewritten using complex memory polynomials as

$$X_{\mathrm{A}}(n) = \sum_{i\geq 0}^{N_i} \mathcal{K}_i\left[E_{\mathrm{P}}^2(n-i)\right] X_{\mathrm{P}}(n-i).$$

The PD output is then rewritten using complex notation as

$$X_{\mathrm{P}}(n) = \sum_{j\geq 0}^{N_j} \mathcal{L}_j(n) X(n-j).$$

On cascading the linearization and the PA, we then have

$$X_{\mathrm{A}}(n) = \sum_{i\geq 0}^{N_i}\sum_{j\geq 0}^{N_j} \mathcal{K}_i\left[E_{\mathrm{P}}^2(n-i)\right] \mathcal{L}_j(n-i) X(n-i-j).$$

Enforcing a linear gain $\alpha G_{\mathrm{A}}^{\mathrm{lin},0}$ for the PA reduces this equation to the following system of equations:

$$\sum_{i\geq \max(0,k-N_j)}^{\min(k,N_i)} \mathcal{K}_i\left[E_{\mathrm{P}}^2(n-i)\right] \mathcal{L}_{k-i}(n-i) = \alpha G_{\mathrm{A}}^{\mathrm{lin},0}\delta_{k,0} \qquad \forall\ 0\leq k\leq N_j+N_i.$$

The final solution for an arbitrary number of taps is then

$$\begin{aligned} \mathcal{L}_0(n) &= \frac{\alpha G_{\mathrm{A}}^{\mathrm{lin},0}}{\mathcal{K}_0\left[E_{\mathrm{P}}^2(n)\right]}, \\ \mathcal{L}_k(n) &= -\frac{1}{\mathcal{K}_0\left[E_{\mathrm{P}}^2(n)\right]} \sum_{i\geq \max(1,k-N_j)}^{\min(k,N_i)} \mathcal{K}_i\left[E_{\mathrm{P}}^2(n-i)\right] \mathcal{L}_{k-i}(n-i) \qquad \forall\ 0\leq k\leq N_j. \end{aligned}$$

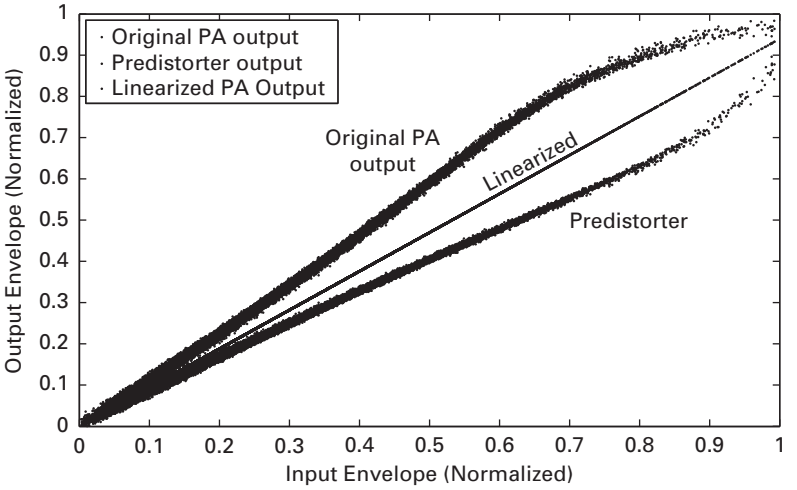


Figure 9.7 An AM/AM plot for an OFDM signal with 8 dB PARP before (dots, top) and after (dots, center) predistortion. The predistorter output (dots, bottom) is also shown.

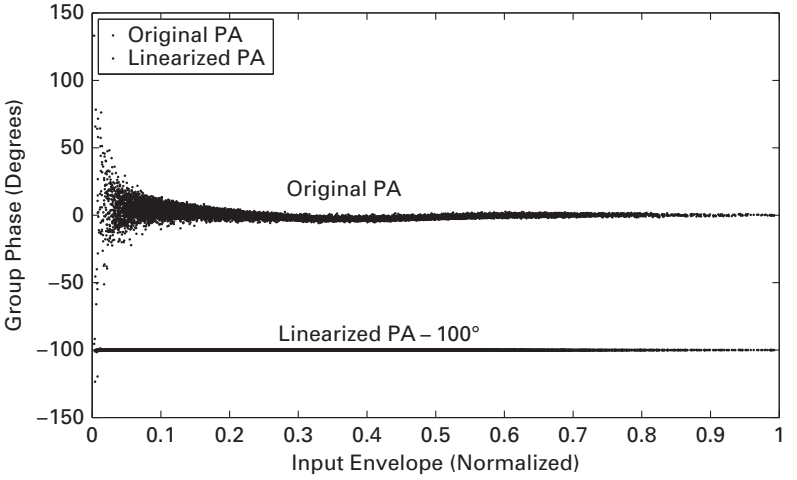


Figure 9.8 An AM/PM plot for an OFDM signal with 8 dB PARP before (dots, top) and after (dots, bottom) predistortion.

This general solution can be verified to reduce to the two-tap case given above. For the two-tap case, the weight of the predistortion correction brought about by the tap k of the predistorter is approximately given by $|\mathcal{K}_1(E_P)/\mathcal{K}_0(E_P)|^k$ in the limit of sufficiently high oversampling. The criterion for convergence is then given by $|\mathcal{K}_1(E_P)/\mathcal{K}_0(E_P)| < 1$.

Simulation results for an OFDM signal are shown in Figures 9.7–9.9. The two-tap PA model is extracted from measurements performed on an 80-W GaN Doherty PA. The signal used for both modeling and linearization consists of a two-carrier 20-MHz

Table 9.1. ACPR of input data and PA output data with and without predistortion

	ACPR for LSB	ACPR for USB
Input data	−48.72	−43.47
Original PA	−32.17	−33.55
Linearized PA	−48.70	−43.46

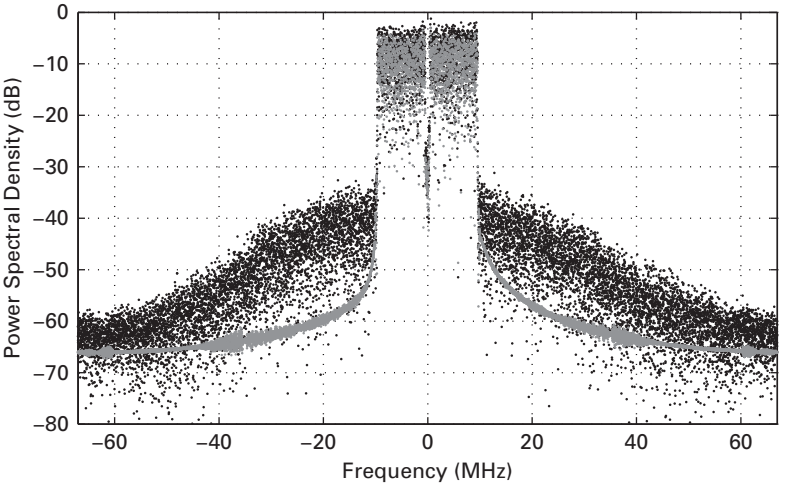


Figure 9.9 Power spectral density for an OFDM signal with 8 dB PARP before (black dots) and after (gray dots) predistortion.

WiMAX signal with 8 dB PARP after signal conditioning with crest-factor reduction (CFR). As is shown in Figures 9.7 and 9.8, the AM/AM and AM/PM curves depart from straight lines, indicating the device nonlinearity. Further, they exhibit clear memory effects in the thickness of the curves indicating that multiple output amplitudes and phases are observed from the same instantaneous input envelope. Note that the predistorter generates an output signal that also exhibits memory effects in amplitude (see Figure 9.7) and phase (not shown) in order to suppress the memory effects of the PA. A predistorter with 11 taps is used to linearize the 2-tap PA. The AM/AM and AM/PM curves for the linearized PA take the form of straight and thin lines, indicating that the amplitude and phase have both been linearized and that memory effects have greatly decreased.

The extent of the linearization is also demonstrated by the power spectral density shown in Figure 9.9. The ACPRs for the LSB and USB are given in Table 9.1. The linearizer is seen to have recovered, up to four digits, the ACPR of the original input two-carrier OFDM signal. The spectral regrowth (ACPR) in this simulation example is therefore limited not by the linearized PA but by the CFR stage which had introduced undesirable spectral regrowth (see Chapter 1) in the process of reducing the PAPR from

10 to 8 dB. Obviously, in a physical system, the actual performance of the linearizer with the original PA will depend on how accurately the memory-polynomial model used is actually predicting the PA performance.

9.4 Two-band frequency-selective predistorter

To conclude this chapter on linearization, we consider now the case of a multi-band PA. We define a multi-band PA as a multi-carrier PA for which the subcarriers are sufficiently separated. If the subcarrier separation is sufficiently large, strong memory effects can be expected from frequency-dependent nonlinearities. To linearize such a PA, a wideband predistorter with a large number of taps may then be required. An alternative approach is to use a frequency-selective algorithm [8] [9] that addresses the linearization of each subcarrier band separately while accounting for their interactions.

Let us consider the case of the two-band PA model discussed in Chapter 6. The cascaded two-band predistorter and two-band power amplifier are shown in Figure 9.10. If we account for $2 \times O$ intermodulation bands, the PA output then reduces to

$$I_{\mathrm{A,U}} = \sum_{k=1}^{O} I'_k, \qquad Q_{\mathrm{A,U}} = \sum_{k=1}^{O} Q'_k, \qquad X_{\mathrm{A,U}} = I_{\mathrm{A,U}} + jQ_{\mathrm{A,U}},$$
$$I_{\mathrm{A,L}} = \sum_{k=-1}^{-O} I'_k, \qquad Q_{\mathrm{A,L}} = \sum_{k=-1}^{-O} Q'_k, \quad X_{\mathrm{A,L}} = I_{\mathrm{A,L}} + jQ_{\mathrm{A,L}},$$

where I'_k and Q'_k are the scaled and phase-shifted versions of the intermodulation terms I_k and Q_k obtained from

$$X'_k = \mathcal{K}_k X_k, \qquad X'_k = I'_k + jQ'_k, \quad X_k = I_k + jQ_k,$$

with

$$\mathcal{K}_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) = G_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) + jH_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2),$$
$$X_{\mathrm{P,U}} = I_{\mathrm{P,U}} + jQ_{\mathrm{P,U}}, \qquad X_{\mathrm{P,L}} = I_{\mathrm{P,L}} + jQ_{\mathrm{P,L}},$$

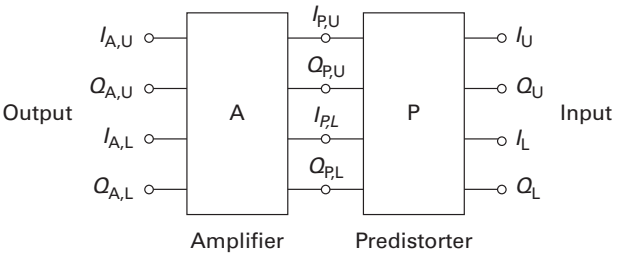


Figure 9.10 A cascaded two-band predistorter and amplifier.

where the intermodulation terms I_k and Q_k themselves are given in Section 6.4.3. For example, in the present notation we have, for k varying from -3 to 3,

$$\begin{aligned}
I_{-3}\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) &= I_{\mathrm{P,U}} I_{\mathrm{P,L}}^2 - I_{\mathrm{P,U}} Q_{\mathrm{P,L}}^2 + 2 Q_{\mathrm{P,U}} I_{\mathrm{P,L}} Q_{\mathrm{P,L}},\\
Q_{-3}\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) &= 2 I_{\mathrm{P,U}} I_{\mathrm{P,L}} Q_{\mathrm{P,L}} - Q_{\mathrm{P,U}} I_{\mathrm{P,L}}^2 + Q_{\mathrm{P,U}} Q_{\mathrm{P,L}}^2,\\
I_{-1} &= I_{\mathrm{P,L}}, \qquad Q_{-1} = Q_{\mathrm{P,L}},\\
I_1 &= I_{\mathrm{P,U}}, \qquad Q_1 = Q_{\mathrm{P,U}},\\
I_3\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) &= I_{\mathrm{P,U}}^2 I_{\mathrm{P,L}} + 2 I_{\mathrm{P,U}} Q_{\mathrm{P,U}} Q_{\mathrm{P,L}} - Q_{\mathrm{P,U}}^2 I_{\mathrm{P,L}},\\
Q_3\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) &= -I_{\mathrm{P,U}}^2 Q_{\mathrm{P,L}} + 2 I_{\mathrm{P,U}} Q_{\mathrm{P,U}} I_{\mathrm{P,L}} + Q_{\mathrm{P,U}}^2 Q_{\mathrm{P,L}}.
\end{aligned}$$

The linearization of the two-band PA therefore leads to the following system of equations:

$$\begin{aligned}
\sum_{k=1}^{O} \mathcal{K}_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) X_k\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) &= \alpha G_{\mathrm{A}}^{\mathrm{lin},0} X_{\mathrm{U}},\\
\sum_{k=-1}^{-O} \mathcal{K}_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) X_k\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) &= \alpha G_{\mathrm{A}}^{\mathrm{lin},0} X_{\mathrm{L}}.
\end{aligned} \tag{9.11}$$

The unknowns are the output signals $X_{\mathrm{P,L}} = I_{\mathrm{P,L}} + jQ_{\mathrm{P,L}}$ and $X_{\mathrm{P,U}} = I_{\mathrm{P,U}} + jQ_{\mathrm{P,U}}$ of the predistorter. The inversion of this two-band PA model leads to a fairly complex system of transcendental equations to solve for $X_{\mathrm{P,L}}$ and $X_{\mathrm{P,U}}$ at each input $X_{\mathrm{L}} = I_{\mathrm{L}} + jQ_{\mathrm{L}}$ and $X_{\mathrm{U}} = I_{\mathrm{U}} + jQ_{\mathrm{U}}$:

$$\begin{aligned}
X_{\mathrm{P,U}} &= \frac{1}{\mathcal{K}_1(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2)}\\
&\quad \times \left[\alpha G_{\mathrm{A}}^{\mathrm{lin},0} X_{\mathrm{U}} - \sum_{k=2}^{O} \mathcal{K}_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) X_k\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) \right],\\
X_{\mathrm{P,L}} &= \frac{1}{\mathcal{K}_{-1}(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2)}\\
&\quad \times \left[\alpha G_{\mathrm{A}}^{\mathrm{lin},0} X_{\mathrm{L}} - \sum_{k=-2}^{-O} \mathcal{K}_k(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) X_k\big(X_{\mathrm{P,L}}, X_{\mathrm{P,U}}\big) \right].
\end{aligned}$$

An improved insight is gained if we seek a perturbative solution. For a third-order Taylor expansion of the $\mathcal{K}_k$ functions using the notation introduced in Figure 6.15, we have

$$\begin{aligned}
\mathcal{K}_1(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) &= H_{1p} + H_{3pm}|X_{\mathrm{P,L}}|^2 + H_{3pp}|X_{\mathrm{P,U}}|^2,\\
\mathcal{K}_{-1}(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) &= H_{1m} + H_{3mm}|X_{\mathrm{P,L}}|^2 + H_{3mp}|X_{\mathrm{P,U}}|^2,\\
\mathcal{K}_3(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) &= H_{3p},\\
\mathcal{K}_{-3}(|X_{\mathrm{P,U}}|^2, |X_{\mathrm{P,L}}|^2) &= H_{3m}.
\end{aligned}$$

Now, using the following perturbative expansion for the unknown predistortion variables $X_{\mathrm{P,L}}$ and $X_{\mathrm{P,U}}$ to determine

$$X_{\mathrm{P,U}} = X_{\mathrm{P,U}}^{(0)} + \Delta X_{\mathrm{P,U}}^{(1)} + \Delta X_{\mathrm{P,U}}^{(3)},$$
$$X_{\mathrm{P,L}} = X_{\mathrm{P,L}}^{(0)} + \Delta X_{\mathrm{P,L}}^{(1)} + \Delta X_{\mathrm{P,L}}^{(3)},$$

the final system of equations to solve reduces to

$$\begin{aligned}
H_{1p} X_{\mathrm{P,U}}^{(0)} &= \alpha G_{\mathrm{A}}^{\mathrm{lin},0} X_{\mathrm{U}}, \\
H_{1m} X_{\mathrm{P,L}}^{(0)} &= \alpha G_{\mathrm{A}}^{\mathrm{lin},0} X_{\mathrm{L}}, \\
H_{1p}\, \Delta X_{\mathrm{P,U}}^{(1)} + \left[H_{3pm} \left| X_{\mathrm{P,L}}^{(0)} \right|^2 + H_{3pp} \left| X_{\mathrm{P,U}}^{(0)} \right|^2 \right] X_{\mathrm{P,U}}^{0} &= 0, \\
H_{1m}\, \Delta X_{\mathrm{P,L}}^{(1)} + \left[H_{3mm} \left| X_{\mathrm{P,L}}^{(0)} \right|^2 + H_{3mp} \left| X_{\mathrm{P,U}}^{(0)} \right|^2 \right] X_{\mathrm{P,U}}^{0} &= 0, \\
H_{1p}\, \Delta X_{\mathrm{P,U}}^{(3)} + H_{3p} X_3\left(X_{\mathrm{P,L}}^{(0)}, X_{\mathrm{P,U}}^{(0)} \right) &= 0, \\
H_{1m}\, \Delta X_{\mathrm{P,L}}^{(-3)} + H_{3m} X_{-3}\left(X_{\mathrm{P,L}}^{(0)}, X_{\mathrm{P,U}}^{(0)} \right) &= 0,
\end{aligned} \tag{9.12}$$

where $G_{\mathrm{A}}^{\mathrm{lin},0} = \min(H_{1p}, H_{1m})$. The solution of this system is therefore

$$\begin{aligned}
X_{\mathrm{P,U}}^{(0)} &= \frac{\alpha G_{\mathrm{A}}^{\mathrm{lin},0}}{H_{1p}} X_{\mathrm{U}}, \\
\Delta X_{\mathrm{P,U}}^{(1)} &= -\left[\frac{H_{3pm}}{H_{1p}} \left| X_{\mathrm{P,L}}^{(0)} \right|^2 + \frac{H_{3pp}}{H_{1p}} \left| X_{\mathrm{P,U}}^{(0)} \right|^2 \right] X_{\mathrm{P,U}}^{0}, \\
\Delta X_{\mathrm{P,U}}^{(3)} &= -\frac{H_{3p}}{H_{1p}} X_3\left(X_{\mathrm{P,L}}^{(0)}, X_{\mathrm{P,U}}^{(0)} \right), \\
X_{\mathrm{P,L}}^{(0)} &= \frac{\alpha G_{\mathrm{A}}^{\mathrm{lin},0}}{H_{1m}} X_{\mathrm{L}}, \\
\Delta X_{\mathrm{P,L}}^{(1)} &= -\left[\frac{H_{3mm}}{H_{1m}} \left| X_{\mathrm{P,L}}^{(0)} \right|^2 + \frac{H_{3mp}}{H_{1m}} \left| X_{\mathrm{P,U}}^{(0)} \right|^2 \right] X_{\mathrm{P,L}}^{0}, \\
\Delta X_{\mathrm{P,L}}^{(3)} &= -\frac{H_{3m}}{H_{1m}} X_{-3}\left(X_{\mathrm{P,L}}^{(0)}, X_{\mathrm{P,U}}^{(0)} \right).
\end{aligned}$$

Six parameters are therefore required in the third-order perturbative approximation to linearize the two-band PA.

The performance of the proposed linearization algorithm was investigated using wideband WCDMA signals. Figures 9.11 and 9.12 show the frequency-selective linearization of a two-carrier WCDMA signal for a 10-W LDMOS PA with 14 dB gain for 34.8 dBm (3 W) output power. The vertical scale is 5 dB per division and the horizontal scale is 15 MHz per division for a total span of 60 MHz. Each WCDMA band has a bandwidth of 5 MHz. The centers of the two bands are separated by 15 MHz. As indicated in Figure 9.12(a), each band is immediately flanked on both sides by spectral-regrowth bands (each of bandwidth about 5 MHz) originating from the interaction of each band with itself via H_{3mm} and H_{3pp} and with the adjacent band via H_{3mp} and

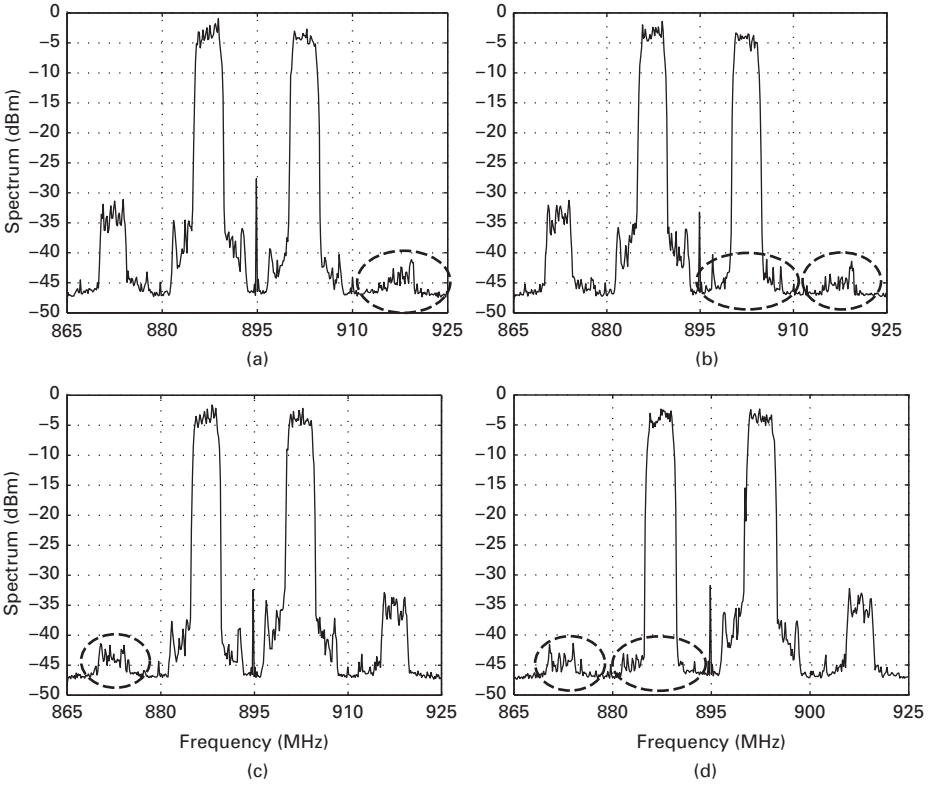


Figure 9.11 Four-band frequency-selective (a) USB interband, (b) USB inband plus interband, (c) LSB interband, and (d) LSB inband plus interband vectorial predistortion linearization of a two-carrier WCDMA signal for a PA with differential memory. (From [8] with permission, ©2008 IEEE.)

H_{3pm}. In addition, the two bands interfere with each other due to the terms H_{3m} and H_{3p} and generate two intermodulation bands at −22.5 MHz and +22.5 MHz relative to the LO (center tone).

The ability of the frequency-selective predistortion-linearization algorithm to reduce the spectral *interband* regrowth of either the LSB only or the USB only is demonstrated in Figures 9.11(a) and (c). Further, as shown in Figures 9.11(b) and (d), not only the intermodulation bands at −22.5 and +22.5 MHz but also the *in-band* spectral regrowth surrounding the original bands themselves can be reduced. The proposed differential algorithm can address separately six types of in-band and interband spectral regrowth in the LSB and USB.

Finally, Figure 9.12(b) shows the combined reduction of the in-band and interband spectral regrowth on both sides of the LO. These results demonstrate the ability of this algorithm to linearize a PA exhibiting differential memory effects ($H_{3m} \neq H_{3p}$) as defined in Section 6.4.1. Higher-order expansion can be used to improve the QML linearization (AM/AM and AM/PM) performance. In addition, memory polynomials can

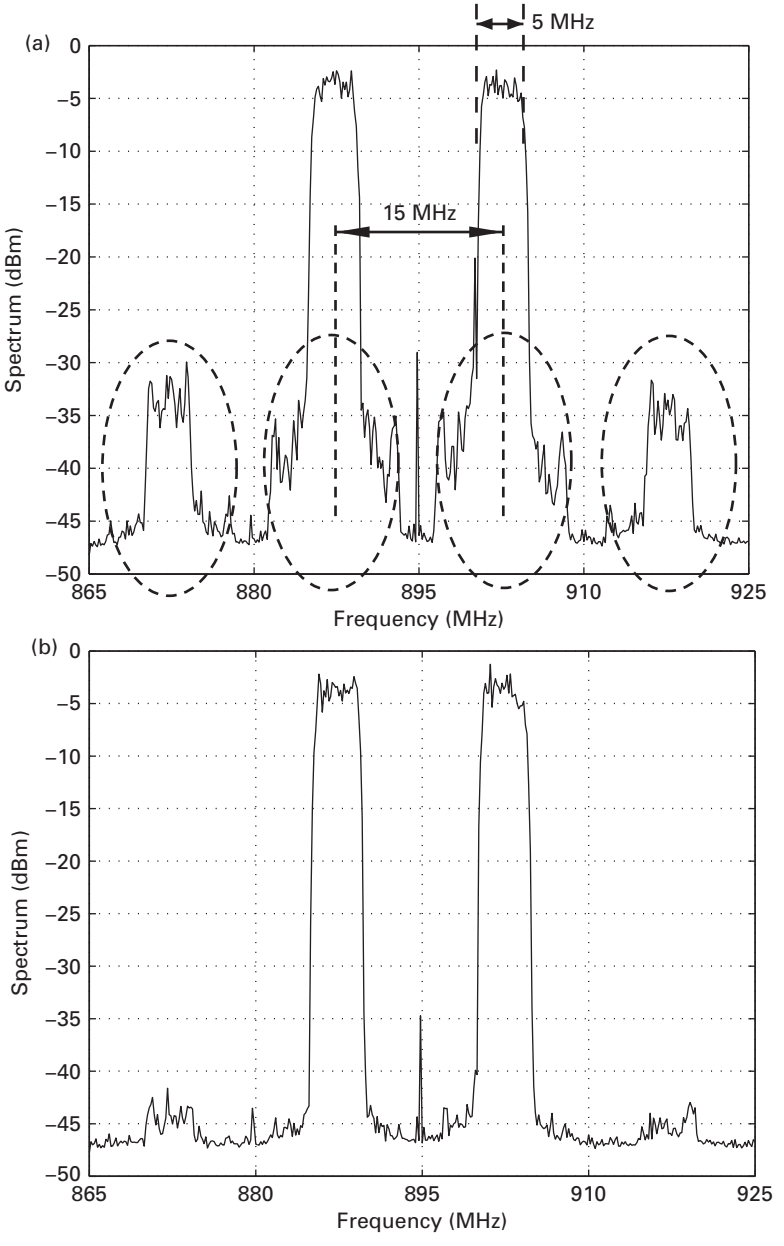


Figure 9.12 (a) Nonlinearized spectrum and (b) joint in-band and interband LSB and USB frequency-selective predistortion linearization of a two-carrier WCDMA signal in a PA with differential memory. (From [8] with permission, ©2008 IEEE.)

be used to account for memory effects. Note that the nonlinearities which extend over a bandwidth of 50 MHz were reduced in this demonstration by using real-time processing [8], with no feedback loop required. However, the implementation of an adaptive algorithm using a feedback path to extract the required linearization coefficients is

greatly facilitated by the frequency-selective nature of the linearization [9]. Indeed, the receiver bandwidth just needs to be centered via a tunable LO on the band to be linearized, since for mild nonlinearities each band is linearized separately.

References

[1] J. S. Kenney, W. Woo, L. Ding, R. Raich, H. Ku, and G. T. Zhou, "The impact of memory effects on predistortion linearization of RF power amplifiers," in *Proceedings of the 8th International Symposium on Microwave and Optical Techniques*, Montreal, pp. 189–193, 2001.

[2] W. Dai and P. Roblin, "Distributed and multi-time-constant electro-thermal modeling and its impact on ACPR in RF predistortion," in *62nd ARFTG Conference Digest*, Denver, CO, 2003.

[3] J. Vuolevi and T. Rahkonen, *Distortion in RF Power Amplifiers*, Artech House, 2003.

[4] S. Boumaiza, J. Gauthier, and F. M. Ghannouchi, "Dynamic electro-thermal behavioral model for RF power amplifiers," in *Digest, 2003 IEEE MTT-S International Microwave Symposium*, Vol. 1, pp. 351–354, 2003.

[5] S. Boumaiza and F. Ghannouchi, "Thermal memory effect modeling and compensation in RF power amplifiers and predistortion linearizer," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 51, No. 12, pp. 2427–2433, Dec. 2003.

[6] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters*, Vol. 37, No. 23, pp. 1417–1418, Dec. 2001.

[7] L. Ding, G. T. Xhou, D. R. Morgan, Z. Ma, J. S. Kenney, J. Kim, and C. R. Giardina, "A robust digital baseband predistorter constructed using memory polynomials," *IEEE Journal of Communication*, Vol. 52, No. 1, pp. 159–165, Jan. 2004.

[8] P. Roblin, S. K. Myoung, D. Chaillot, Y. G. Kim, A. Fathimulla, J. Strahler, and S. Bibyk, "Frequency selective predistortion linearization of RF power amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 56, No. 1, pp. 65–76, Jan. 2008.

[9] Xi Yang, P. Roblin, D. Chaillot, S. Mutha, J. Strahler, J. Kim, M. Ismail, J. Wood, and J. Volakis, "Fully orthogonal multi-carrier predistortion linearization for RF power amplifiers," in *IEEE MTT-S International Microwave Symposium Digest, MTT '09*, pp. 1077–1080, 2009.

Index