RF CMOS CLASS C
POWER AMPLIFIERS FOR
WIRELESS COMMUNICATIONS

by

Ramakrishna Sekhar Narayanaswami

Memorandum No. UCB/ERL M01/37

12 December 2001

cover

RF CMOS CLASS C
POWER AMPLIFIERS FOR
WIRELESS COMMUNICATIONS

by

Ramakrishna Sekhar Narayanaswami

Memorandum No. UCB/ERL M01/37

12 December 2001

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

RF CMOS Class C Power Amplifiers for Wireless Communications

by

Ramakrishna Sekhar Narayanaswami

B.S. (University of California, Berkeley) 1993
M.S. (University of California, Berkeley) 1998

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy
in

Engineering - Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Paul R. Gray, Chair
Professor Robert G. Meyer
Professor J. Karl Hedrick

Fall 2001

The dissertation of Ramakrishna Sekhar Narayanaswami is approved:

Paul R. Gray Nov 4, 2001

Chair Date

Robert G Meyer Nov. 6 2001

Date

J Karl Hedrick NOV. 8 2001

Date

University of California, Berkeley

Fall 2001

RF CMOS Class C Power Amplifiers for Wireless Communications

by

Ramakrishna Sekhar Narayanaswami

Abstract

RF CMOS Class C Power Amplifiers for Wireless Communications

by
Ramakrishna Sekhar Narayanaswami

Doctor of Philosophy in Engineering-Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Paul R. Gray, Chair

Recent efforts in the design of integrated circuits for RF communication transceivers have focussed on achieving higher levels of integration by including more and more analog functional blocks onto a single silicon CMOS chip. One of the final blocks that has yet to be successfully integrated is the Power Amplifier (PA). The PA is the final functional block in the transmit path; its function is to amplify the signal to be transmitted to the required transmit power level. In general, PAs are difficult to integrate in CMOS because of technology limitations that severely limit the efficiency of the PA.

This thesis describes theoretical analysis and circuit techniques for the design and implementation of RF Class C PAs in CMOS technologies. To date, very few methods exist for designing Class C PAs; in the past, much of the design process has been empirical. The theoretical work in this thesis attempts to describe a method for designing a Class C PA in CMOS without resorting to blind use of a circuit simulator. Using fourier series analysis, the drain current waveform of a CMOS Class C PA (which is dependent both on the input and output voltage waveforms) is determined to first-order in order to generate an approximate solution to the design goal before a circuit analysis tool is required. Circuit techniques used to combat the technology limitations imposed by CMOS technologies include the use of differential circuits in the signal path, cascoded stages and a modified tuning method which allowed for the use of extremely large output devices but not requiring passive devices that were not feasible in a CMOS technology.

A 1.75-GHz, 1.7W CMOS PA was designed in an STMicroelectronics 0.35-µm, five-metal doubly-poly CMOS process and implemented both as a stand-alone chip and as part of a fully-integrated transmitter chip. In simulation, the peak efficiency of the PA was 40%. Due to an over-estimation of the quality factors of the on-chip spiral inductors used in the design, the evaluation of the PA revealed a peak power of just over 1-W and a peak efficiency of 27%. The PA did meet the spectral mask requirements of the DCS1800 cellular communications system for which it was designed, and comparison of the PA output of the integrated version indicated that the PA amplified the desird signal without too much degradation of the signal.

Paul R. Gray

Paul R. Gray, Chairman of Committee

Table of Contents

Acknowledgments

When starting the long road towards obtaining a Ph.D., I was told that getting an M.S. (at least at a university that requires a research project for the M.S.) would give one an idea of what doing research was like; it was not necessary to get a Ph.D. for that purpose. Rather, the Ph.D. was more a labor of love than an introduction to research, and that one should be well aware of that before embarking on such a long journey. More than eight long years later, I can say without hesitation that the above is true, and, as such a labor, this dissertation and all the work that went into it would not have been possible without the assistance and support of a great many people.

First and foremost, I would like to thank my advisor, Professor Paul Gray, for his support over the years. His advice on technical matters was invaluable, and his guidance both on my individual project as well as the larger GSM transceiver project was critical to the project's success. I would also like to thank Professor Robert Meyer for his assistance along the way, as his knowledge and experience were also extremely valuable. Along with Professors Gray and Meyer, I would also like the thank Professor Seth Sanders and Professor Karl Hedrick for serving on my qualifying exam committee. Further, I would like to thank Professors Gray, Meyer, and Hedrick for reading this dissertation.

In any endeavor of this magnitude, the work at hand could definitely not have been completed without significant input and assistance from my fellow students. Jeff Weldon, who coordinated much of the work on the transmit side of the GSM transceiver project, was a great friend and colleague throughout this work, and of significant assistance along the way. I don't know that I can ever forget the countless hours required to put together the transmitter and then test it once it had been fabricated. Jeff's help in both discussing technical issues as well as keeping things light with discussions about sports and other "less-critical" topics was invaluable. Others

involved with the project who were extremely helpful include Chris Rudell, Jeff Ou, Li Lin, Martin Tsai, and Luns Tee, as well as the visiting industrial fellows Masa Otsuka, Danilo Gerna, and Sebastien Dedieu.

More generally, my colleagues in Professor Gray's research group were extremely helpful as well. Those mentioned in the previous paragraph were definitely of great assistance, but a few others need to be mentioned here. First, my cube-mates from cube 550-C5 were a great group of guys to have as cube-mates. Andy Abo, Keith Onodera, and Jeff Ou made the long days and nights of work in Cory Hall pass by much more quickly, and offered much relief from the tedium of graduate school through long discussions on Cal Athletics and life in general as well as far too many meal and coffee breaks (and poker nights!). Jeff Ou's insights into RF modelling and devices in general were especially helpful. And Keith and Andy always had their own technical insights that they were willing to offer when I needed it. Overall, I couldn't have asked for a better group of cubemates.

Earlier in my career, the senior graduate students of the time, including Ken Nishimura, Greg Uehara, Cormac Conroy, Robert Neff, Thomas Cho, Todd Weigandt, and Dave Cline were kind enough to impart some of their experience and wisdom to me. Those who joined the group with me as well as newer students who joined later were always ready to discuss topics both technical and non-technical, and their contributions were invaluable. Srenik Mehta, Arnold Feldman, George Chien, Carol Barrett, and Troy Robinson were always available for discussions and exchanging ideas, whether it was circuits or exchanging ideas about our favorite science fiction and fantasy authors. Finally, in the later years, several newer graduate students have entered the group, and I hope I was able to be as helpful to them as some of the senior graduate students were to me when I started, and to wish them good luck in their careers. They include Ryan Bocock, Yun Chiu, Cheol-Woong Lee, and Tim Wongkomet.

Several graduate students in other research groups were also very helpful along the way. Tom Burd and Dennis Yee deserve great thanks, for being great friends as well as helping technically. Ali Niknejad and Manolis Terrovitis from Professor Meyer's group were always available to discuss technical concerns, as were the RF group from Professor Brodersen's research group: Chinh Doan, Dave Sobel, Brian Limketkai, Sayf Alalusi, Johan Vanderhaegen, and Dennis Yee.

I'd also like to recognize a couple of graduate students who were not in my technical field, but definitely added to the overall experience of my time here at Cal. They include Adrian Isles, who has been a good friend ever since I convinced him to come to Cal (isn't that right, Adrian?), and Sunil Khatri, who increased my appreciation for the blues and who helped me with my attempts to retain the Hindi I learned.

Any discussion of my time here at Cal (at least as a graduate student) would most certainly not be complete without acknowledging 920 Keeler, the house where I lived for most of my tenure as a graduate student, and my housemates in that house. Special thanks must go to all the people who graced that house, as all my housemates have become good friends and provided me with some great memories along the way. Andy Abo was not only my cubemate but my housemate, and I will always remember his love for video games and his enjoyment at talking back to the TV, as well as his "textbook" basketball jump shot (which was indeed out of a textbook!). I will also remember his "curious" means for obtaining a Costco card for me. I will always remember Srenik Mehta living in the kitchen for a month, Tony Stratakos' joy for life (and his numerous snapple bottles), and Chris Rudell's 30th birthday party, which was definitely the party to end all parties. I will also always remember I have lived in the same house as Chris longer than anyone outside my immediate family; the fact that I was able to live in the same house as these guys for such a long time is a credit to our friendships and the caliber of their character. I can't continue without mentioning the my newer housemates: Joe Seeger, who first introduced the concept of housemates having schedules that were totally out of phase with each other and with whom I had many good technical discussions, Justin Black, who took the concept of schedules out of phase to the next level, and my newest roommate, Shyam Lakshmin, who has infused the house with his youthful energy (and his love for all things FSU). All my housemates are deserving of a great deal of my appreciation for making my home life a great joy and making the path to my Ph.D. that much easier.

It is certainly not feasible to thank all my friends in this space, but a few of the more important friends who helped me throughout my Ph.D. career should be acknowledged here. Vikram and Dipanwita Amar have been exceptional friends and deserve more thanks than I can express here. Lapoe Lynn, Kevin Stone, Jennifer Cheng, and Manish Mehta, all friends from my undergraduate days (and some from my early graduate days), have also been great friends, and also have my greatest appreciation.

Finally, it is absolutely true that none of this Ph.D. would have been possible without the love and support of my family: my parents and my sister, Kala. Their love and overwhelming support in the face of any obstacle or adversity I faced was not only instrumental but essential in the completion of this work. I would not be where I am today without their contributions and the foundation that they have provided throughout the years.

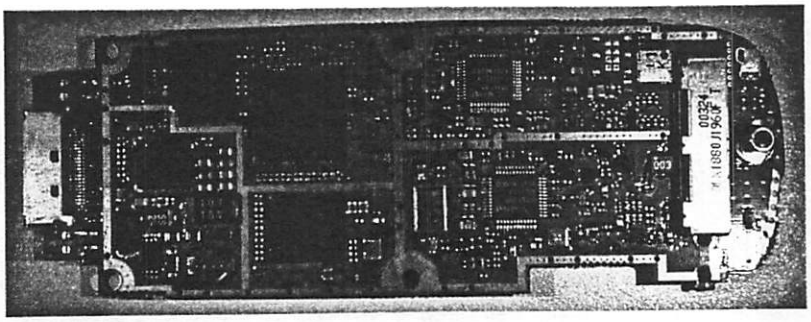
Funding for this work was provided by the National Defense Science and Engineering Graduate Fellowship, as well as the National Science Foundation. Fabrication facilities and support were provided by STMicroelectronics.

1 Introduction

1.1 Background

In recent years, the level of interest in wireless communication links and the devices which can support those links has exploded[1]. New standards are being both approved and designed in order to tap into the exploding market; many of these new standards attempt to connect devices and/or appliances in the home using lower-performance radio transceivers. In addition, the desire for Internet connectivity using cellular phones or Personal Digital Assistants (PDAs), i.e. Palm or Windows-CE based handheld devices, has dramatically increased the demand for universal wireless connectivity. In order for these standards and the companies that support them to stay competitive, low-cost, small form-factor portable wireless communications devices are a critical component; in order to attract a large group of users, cost is a very important concern. As such, the desire for a low-cost wireless device with reasonable performance has exploded as well.

Current implementations of wireless communications devices, such as cellular phones or cordless phones, employ several chips implemented in different semiconductor technologies in order to implement the analog radio-frequency (RF) components and lower frequency components as well as the digital components. As an example, photos of the inner workings of two cellular phones are shown in Figure 1-1..



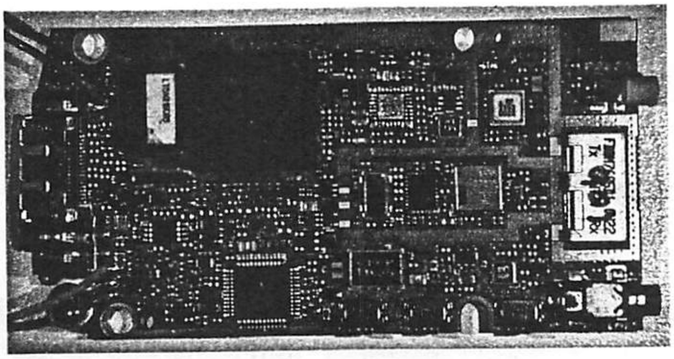


Figure 1-1. Cellular Phones

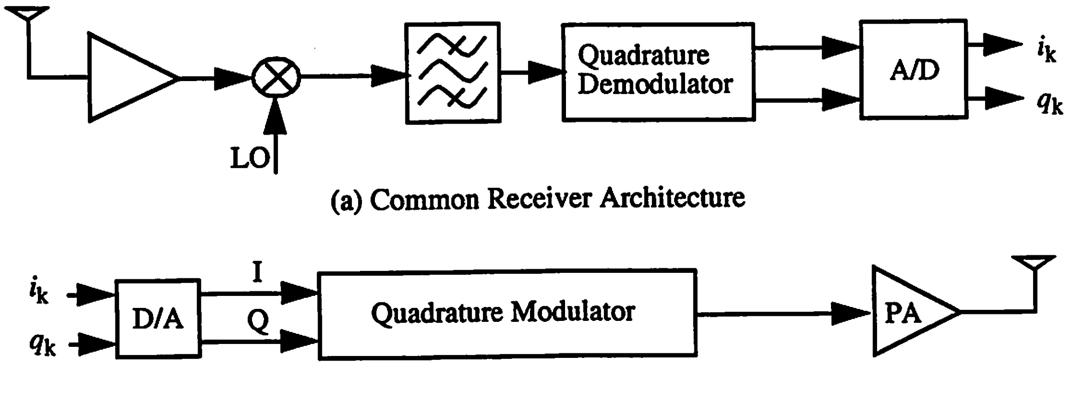
The two photos show that each of these phones is comprised of several distinct chips as well as a large number of discrete components (often passive components). Often times the chips included in the analog section of the transceiver will include separate chips for the radio frequency (RF) functions (usually at least two distinct chips, if not more), the intermediate frequency (IF) functions, and the baseband analog sections. The blocks required to convert the analog signals to digital signals (and vice versa) may also be separate from the sections listed above, so there may be as many as five chips or more required to perform the functions of the analog functionality from RF to baseband. Furthermore, these chips are normally designed in different fabrication and device technologies, where each of the chips is designed in a technology which is best suited for the particular function being implemented. This multi-chip solution to implementing these portable wireless devices limits the minimum cost and size of the final devices. Further adding to the cost and size is the fact that passing signals between chips is generally more complex than passing signals around on a single-chip, especially at radio frequencies. Issues of matching and driving signals between chips add to the number of discrete components required and also increase the power consumption of the overall implementation, reducing the performance of the overall device. The current implementations, while they are becoming smaller, are still extremely complex and not conducive to significant reductions in cost and size without a radical change in thinking with regards to the implementation of the transceiver.

To that end, the amount of investigation into the feasibility of using CMOS processes in the implementation of the circuitry used in these devices has dramatically increased as well. While CMOS processes are already used for many, if not all, of the lower frequency and digital functions, their acceptance as a reliable and feasible alternative on which to build the high-frequency high-performance analog blocks has been limited. In general, CMOS is an extremely poor technology for high-frequency, high-performance analog functions. For reasons that will be detailed in this work, and have been detailed in others, CMOS is much better suited to low-frequency analog implementations or digital functions; a CMOS transistor models an ideal switch very well, which is extremely useful for digital functions or switched-capacitor analog functions, but it in general has poor current drive and large associated parasitic capacitances, reducing its usefulness in higher-frequency analog functions.

However, what CMOS does provide is an extremely low-cost technology in which to implement circuit blocks, along with the added benefit of potential large-scale integration. Because the baseband analog and digital chips are generally CMOS chips, building the high-frequency, high-performance chips in CMOS would allow those functions to be included in comprehensive, single-chip solutions. Currently, most commercially-available wireless solutions are multi-chip implementations, in which several different technologies, including silicon CMOS (Si-CMOS), silicon bipolar, Gallium Arsenide (GaAs), and/or Silicon Germanium (SiGe). In general, the higher-frequency blocks are done in one of the listed technologies, with the exception of CMOS. A CMOS implementation of the high-frequency analog section would facilitate the integration of all of the radio functions (including high-frequency analog, baseband analog, A/D conversion, and DSP) onto a single-chip.

1.2 Transceiver Architecture

In order to accomplish this, there must be an understanding of what functions must be implemented in the high-frequency analog portion of the RF transceiver, or radio. Figure 1-2. shows the basic architecture of a radio transceiver, which contains two main functions: that of the receiver and that of the transmitter. The function of the analog portion of the receive chain is to take a radio-frequency (RF) analog signal and convert that into the digital bitstream which it represents. The basic blocks in the receive path are a low-noise amplifier (so called because it is critical to amplify the incoming signal without contributing any significant noise), a frequency translation block (generally also known as a mixer), and a block which demodulates the signal and



(a) Common Receiver Architecture

(b) Common Transmitter Architecture

Figure 1-2. Radio Transceiver Architectures

outputs the desired digital bits. A block that is required but which is not directly in the signal path is the local oscillator (LO), which is used by the mixer to frequency translate the signal to low-frequency.

On the transmit side, essentially the reverse path is followed. First, the digital bitstream is modulated and converted to an analog signal. The resulting analog signal is frequency translated by an up-conversion mixer, and the final signal is amplified to the power level that the antenna needs to radiate, as required by the standard. This final block is generally known as a Power Amplifier (PA).

1.3 Power Amplifiers

In RF transceivers, the PA takes a small-amplitude signal at the output RF frequency as its input and drives a high power representation of the input into a low-impedance load (generally the antenna, which is nominally 50Ω). Because the peak power levels required will be significantly higher than the mixer can supply on its own, a separate functional block is required. So the key purpose of the block is high-frequency amplification; as mentioned earlier, CMOS is not ideally suited to this function. While many of the other functional blocks have been implemented in CMOS reliably, the PA is one of the final blocks to be implemented and integrated with the other blocks onto a single CMOS chip.

One of the key reasons for the general industry-wide reluctance to move the implementation of the PA to CMOS is the fact that while the PA is on, it can dominate the power consumption of the entire transceiver. Since most of these portable wireless devices will be powered by a battery with a finite energy reserve, the amount of power is consumed by any given block can be a critical factor in determining the usability of a particular part. As a result, optimizing the power-performance of the PA has been a far more important goal than that of integrating the PA into a transceiver. While the CMOS realization of the PA might reduce the cost of the PA itself, if it significantly reduces the battery-life of the wireless device, there is really no gain (and there might be a potential loss) in that implementation.

Although the implementation of the PA in CMOS has not reached reliable production levels yet, there is a great deal of ongoing research into making the CMOS PA a realizable goal. Why? Because the cost and form-factor benefits from a high-performance CMOS PA are dramatic. If one can avoid having to use a GaAs or SiGe PA in favor of a CMOS PA, there is an inherent cost reduction in that switch. Second, the potential for including that PA in a single-chip CMOS transceiver, or in fact, in a single-chip that would serve as the entire analog *and* digital processing unit for a cellular phone would dramatically reduce the cost and form factor of the PA, allowing for the placing of cellular phone transceivers/processors in almost anything, including something as small as a watch.

1.4 Research Goals

The focus of this research, and this dissertation, is the realization of a high-performance, integrated PA in CMOS. Through the use of PA architecture, or class, as well as circuit design techniques, this work will try to show the feasibility of a CMOS PA implementation, as well as its usefulness in an integrated transceiver. More generally, a key goal of this research is the development of a design methodology for the design of Class-C PAs in CMOS technologies. A few of the key results of this research are listed here:

- A simplified design methodology has been developed for the design of Class-C PAs in CMOS technologies. The design methodology uses a fourier series analysis of the drain current along with an iterative procedure in order to predict the drain current of Class-C PA in which the device operates in all three MOS regions of operation.

- Several different circuit techniques were used in order to overcome the limitations inherent to CMOS technologies, such as limited voltage swing and poor transconductance. New techniques, including the use of cascode inductors and the use of a modified tuning method in order to peak the impedance and therefore the gain at the input to the final drive stages, allowed the use of extremely large devices which facilitated the high output power required for the prototype designed in this work.

- A 1.7-W, 1.75-GHz PA has been designed in this work in a 0.35–μm CMOS process, using a Class-C architecture. Experimental results indicated that the PA actually delivered 1-W of output power to the load, due to lower-than-expected Qs from the on-chip spiral inductors.

- The PA was also integrated into a single-chip CMOS transmitter that implemented all the transmitter blocks from the digital-to-analog converters through the PA as well as two frequency synthesizers. The PA in that implementation also suffered from lower than expected gain due to the lower-than-expected spiral inductor Qs; however, the PA was shown to have a minimal impact on the transmitted signal when compared with the output of the transmitter prior to the PA (after the mixers).

1.5 Thesis Organization

This dissertation is organized as follows. In the next chapter, background information on both transmitters and PAs will be covered. The discussion on transmitters will briefly discuss a new transmitter architecture that allows the use and integration of a non-linear PA; the PA discussion includes PA uses, different classes, key metrics, and an in-depth background on Class C PAs. Chapter 3 will cover a theoretical approach to designing Class C PAs in CMOS, which has previously been missing. Chapter 4 will present an analysis of the limitations of CMOS processes with respect to the implementation of PAs. Chapter 5 will cover the techniques used in this work to combat the limitations described in Chapter 4. Chapter 6 will describe the PA that was included in the integrated transceiver, as well as the corresponding single-chip experimental prototype, and present some simulation results. Chapter 7 will cover the experimental results and performance of the single-chip PA prototype as well as some results from the integrated transceiver. Furthermore, Chapter 7 will also include some discussion on the deviation of the simulated and experimental results and the causes for that deviation. Finally, some conclusions drawn from this effort will be presented in Chapter 8.

2

Power Amplifier Background

2.1 Introduction

In this chapter, a little background information will be provided, both on transmitters and on power amplifiers themselves. First, in order to be able to integrate a non-linear PA in a transmitter that attempts to satisfy a cellular standard, a transmitter architecture that is amenable to PA integration must be used. While the architecture advancements are not apart of the effort detailed in this dissertation but the work of another student[3], the ability to integrate a non-linear PA is contingent on the use of an architecture such as the one described in the next section. Because it is a necessary condition to the use of this non-linear PA, a basic explanation of the concepts is presented here.

The rest of the chapter will discuss PAs in further detail. A basic background on PAs and their function will first be given, followed by a brief introduction to the different classes of PA. Finally, the final section will fully investigate what is known about Class-C PAs, and provide an idea of what else needs to be done to implement a Class-C PA today.

2.2 Transmitter Basics

In today's world, the end goal for many research efforts in implementing low-cost wireless transceivers is the single-chip radio. Currently, most two-way radios, especially those designed for high-performance wireless systems like cellular phone systems, are implemented with multi-chip solutions. The RF, IF, and baseband sections may be comprised of different chips in different technologies, and there may even be further subdivisions within those groups. In order to implement the single-chip radio, a single technology must be chosen, and architectures that enable the integration of all the different blocks onto a single chip must be used.

On the receiver side, there has already been significant research into such architectures, including well-known architectures such as direct-conversion[4]. Newer architectures have also found significant interest, including Low-IF[5] and Wideband-IF-with-Double-Conversion (WIFDC)[6]. On the transmitter side, the level of interest has lagged behind that of the receiver side; however, more and more work is being focused on the transmitter.

One of the key issues in transmitter integration is that significant amounts of filtering are generally needed in the signal path in order to ensure that the output of the PA is free of both spurs and spectral regrowth that would violate either the spectral mask or the level of spurs which a given standard specifies. This is especially true if the PA to be used in the transmit chain is a non-linear PA, in which the level of distortion in the PA input will generally be amplified. If two simple transmitter architectures are examined, it will be apparent that neither is particularly well-suited to full-scale integration. The first of these architectures is the transmitter implementation of the direct-conversion architecture; a block diagram is shown in Figure 2-1. In the

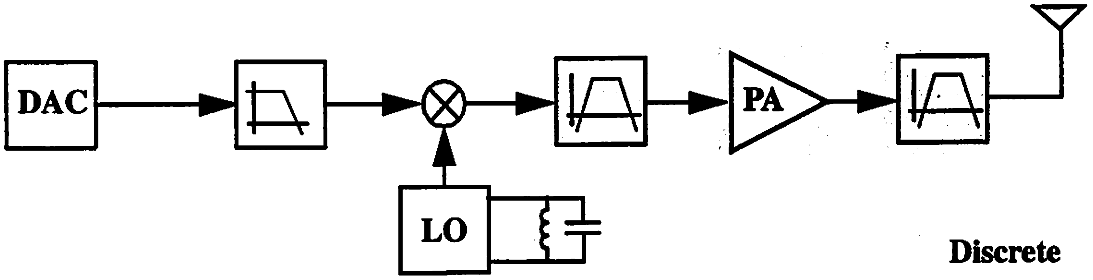


Figure 2-1. Direct Conversion Transmitter

direct-conversion transmitter,, the baseband digital bits representing the in-phase and quadrature channels are converted to analog signals in a digital-to-analog converter (DAC), filtered, then upconverted directly to the RF carrier frequency in one frequency-translation step. In general, at this stage, the signal needs to be filtered before passing through the PA. This is for two reasons: first, the noise outside the transmit band must be reduced, which reduces the filtering required after the PA, and second, third-order intermodulation between frequency components at the input of the PA could mix and sit on top of the desired signal. This filter is difficult to integrate since it must have a narrow pass-band and a large amount of rejection, which cannot be achieved with the low quality-factor (Q) passive components available in a silicon CMOS technology. Furthermore, the frequency synthesizer used to generate the LO frequency needs to be a tunable frequency synthesizer which can tune between a large number of steps which are several orders of magnitude below the center frequency. As a result, the frequency synthesizer will need a high quality-factor LC tank in order to generate a clean LO signal (i.e. one with very low phase-noise); like the filter, this high-Q tank will not be able to be integrated.

A second popular architecture is the transmitter version of the super-heterodyne receiver, which can be thought of as a transmitter using two steps to convert the frequency from baseband to the RF carrier frequency. The block diagram of this architecture is shown in Figure 2-2. In this architecture, once the digital bits are

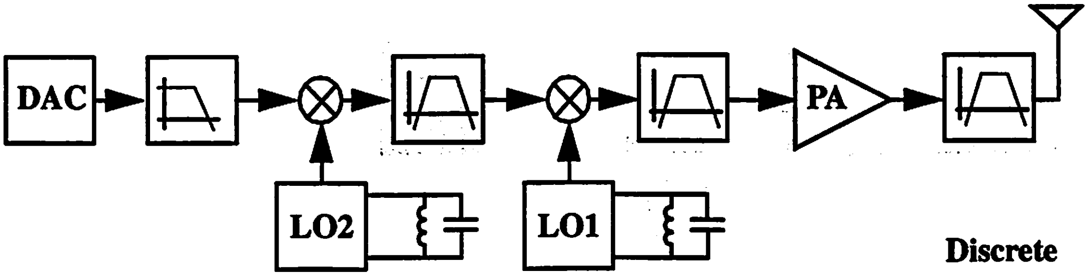


Figure 2-2. Two-Step Transmitter

converted to analog signals in the DACs, the signal is frequency-translated to a fixed intermediate frequency (IF). There, the signal is filtered by a narrow-band filter, which serves the purpose of removing the harmonics of the lower frequency local oscillator (LO2). The LO2 frequency is usually relatively low, when compared with the RF carrier frequency, and as a result, the harmonics of LO2 will cause intermodulation distortion in both the mixers and the PA which can give rise to significant errors in the transmitted signal. Because the LO2 frequency is generally low, the filter will have to have a narrow pass-band as well as lots of suppression in the stop-band, necessitating a high-Q filter. This IF filter is normally implemented as a discrete component. Once the signal has been filtered at IF, it is then frequency translated to the RF carrier frequency. Again, the high frequency local oscillator used in this second mixing (LO1) is the one used to select the channel, and since it must be able to select between a number of closely-spaced channels, the LC tank required for this synthesizer will again likely be discrete. Finally, a pre-PA filter is, like the direct-conversion case, required in the signal path in order to reject the noise and potential for intermodulation. Finally, the signal is passed through the PA. As in the direct-conversion case, the LC tank used in the frequency synthesizer and the pre-PA filter need to be discrete; furthermore, in this super-heterodyne implementation, another discrete element, the IF filter, is added to the signal path. If anything, this architecture is even less amenable to integrating the entire signal path.

An architecture that attempts to solve these problems was presented recently[2] and will be described in more detail shortly[3], and is shown in Figure 2-3..

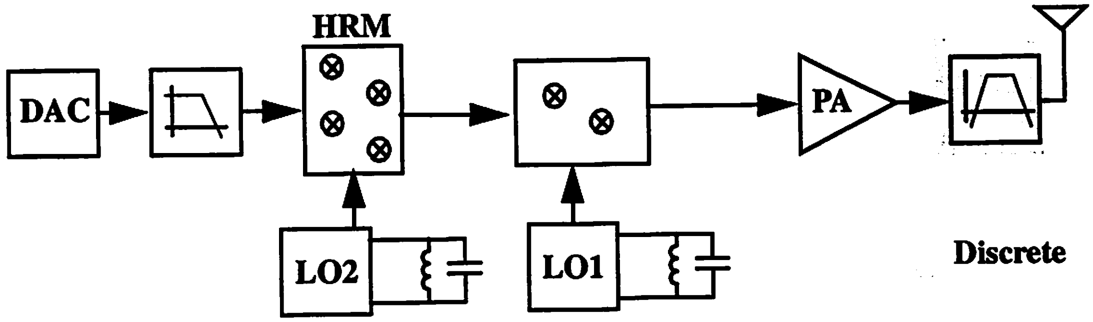


Figure 2-3. Harmonic Rejection Transmitter

This architecture, known as the Harmonic Rejection Transmitter (HRT), includes a special set of mixers known as Harmonic Rejection Mixers (HRM). The HRT architecture allows for the elimination of all of the discrete components required in the previously described architectures, except for the post-PA RF filter. As a result, the entire signal path from the DACs to the output of the PA can be included on a single-chip. The two key innovations in the transmitter include the use of the HRMs and the roles of the frequency synthesizers used. The HRT performs the frequency translation from baseband to the RF carrier frequency in two steps. However, unlike the super-heterodyne transmitter mentioned previously, the roles of the local oscillators are swapped; the high-frequency LO (LO1) is nominally fixed in frequency, and the lower frequency LO (LO2) is the tunable, channel-select oscillator. The benefit of this arrangement is that it allows the use of a wide-bandwidth PLL in LO1, which can be designed to provide low phase noise even if a low-Q LC tank is used[7]. Furthermore, because the channel-selection is performed at a lower frequency, the output of LO2 will inherently have lower phase noise, again allowing the use of low-Q on-chip components. Thus the need for the high-Q discrete components required in the VCOs have been eliminated. Furthermore, the use of the HRMs helps to eliminate the need for the filters used in the signal path before the PA. In standard active mixers (such as the Gilbert Cell-style mixers), the LO signal is applied to the mixers as a square wave. As is well-known, the square wave has large frequency components at the third and fifth harmonics of the fundamental frequency. In the super-heterodyne transmitter, the harmonics of the LO input mix with the baseband signal in the first mixing step. The IF filter is required in the super-heterodyne transmitter in order to prevent those harmonic components from producing distortion in the high-frequency mixers or the PA. In the Harmonic-Rejection Mixers (HRM), however, the LO input is applied not as a square-wave but as a staircase function. The staircase function can be thought of as a three-bit, amplitude-quantized sine wave (similarly, the square-wave can be thought of as a one-bit amplitude-quantized sine wave). This staircase function has no third or fifth harmonic component, and thus there will be no mixing products at those frequencies after the first mixing stage. If the LO2 frequency is low (on the order of 10 MHz or less), the seventh harmonic may still be problematic, as it would only be 70 MHz away, and may still produce intermodulation distortion that mixes with the desired signal. However, if the frequency plan is designed to use a higher LO2 frequency, the harmonic products existing after the first mixing stage would not exist at frequencies which may distort the output signal, and thus the need for the IF filter is eliminated. Furthermore, with the low phase-noise of LO1 and the removal of the harmonic components from the HRM, the pre-PA filter is also not required, and thus the entire signal-path from beginning to end can be integrated if the Harmonic-Rejection Transmitter is used. For more detail on this architecture, the references [2] and [3] should be consulted.

The HRT architecture facilitates the use of a non-linear power amplifier (PA) in the integrated transmitter, and furthermore facilitates the integration of the entire transmit path, from the DACs all the way through the PA, including the frequency synthesizers used. In the next section, PAs will be discussed in significantly more detail, in order to provide a basis from upon which the remainder of this work will be built.

2.3 Power Amplifier Basics

Power Amplifiers (PAs) are used in the transmit chain of communications devices, in order to amplify the signal to the desired power level. That power level is determined by the communications system; it must be high enough such that the amount of power that the receiver is able to sense (taking into account the losses in the communication medium) is adequate to recover the desired signal. For different applications, the order of magnitude of the transmitted power can very greatly. For base-stations used in cellular systems, this can be on the order of tens to hundreds of watts. For satellite communications, this can be on the order of thousands of watts. For portable wireless communications devices, the peak transmitted power is often significantly less; it will vary from tens to hundreds of milliwatts in cordless systems to hundreds of milliwatts to a few watts in cellular systems. Finally, in many of the emerging standards for wireless connectivity in the home (such as Bluetooth or HomeRF), the power will vary between tens of milliwatts and hundreds of milliwatts. When speaking of the power output of these blocks, one common unit used is dBm, which is the output power in dB referenced to 1-mW. That is, the output power in dBm is given by

$$P_{dBm} = 10log\frac{P}{0.001W}, \qquad \text{Eq. 2-1}$$

where P is defined in watts. Thus 1-W is equivalent to 30dBm, 100-mW is equivalent to 20 dBm, etc.

In the cases where the power output is on the order of hundreds of milliwatts or more, the power that the PA needs to deliver to its load in itself is a large percentage of the total power consumed by the entire transmitter. The power that needs to be delivered will be taken from the source that powers the PA; in the case of a portable unit, this will be the battery. In essence, the PA converts the DC power from the battery into RF power delivered to the load. Unless that power conversion is lossless, which is possible only as an ideal abstraction, the PA itself will consume power, over and above what it delivers.

The measure of how much power a PA consumes in this conversion is one of the key performance parameters used to gauge PAs, especially those PAs used for portable applications. Because PAs in portable applications are driven from a source with a finite amount of available energy, power consumed in the PA directly goes to reducing the battery life. This metric is known as the PA's *efficiency*, given by

$$\eta = efficiency = \frac{Power\ Delivered\ to\ Load}{Power\ Drawn\ from\ Supply}. \qquad \text{Eq. 2-2}$$

Since the PA is really converting the DC power of the supply into the RF power delivered to the load, the maximum efficiency is 1, or 100%. That is, if there is no power consumed in the PA - all the power from the supply is sent to the load - both the numerator and the denominator in Eq. 2-2 are the same. However, since this is only ideally possible, the biggest issue in PAs today is maximizing this metric.

Furthermore, there are variations on this metric that give us more information about the PA. The *drain efficiency* is defined as

$$\eta_D = drain\ efficiency = \frac{Power\ Delivered\ to\ Load}{Power\ Consumed\ in\ Final\ Stage}. \qquad \text{Eq. 2-3}$$

This tells us how efficient the final stage, often referred to as the *power stage*, is. The most common efficiency metric used, though, is the ***Power-Added Efficiency*** (PAE). The one thing missing in the previous two versions of efficiency is any idea of how much power is needed to drive the input to the PA. In a real-world PA implementation, the power needed to drive the PA is also drained from the power source, and while that power loss can be accounted for in the previous stage, including it in the PA performance metric allows for two different PAs to be compared. If one PA has a better PAE for the same output power level and same overall efficiency than a second PA, it would be preferable to use the first PA; overall, less power would be drawn from the power source. PAE is defined as

$$PAE = \frac{P_{RF_{OUT}} - P_{RF_{IN}}}{P_{DC}} \qquad \text{Eq. 2-4}$$

So the PA that delivers 1W to the output, consumes 2W, and must be driven by a 100mW input (PAE=45%) is not as good as the PA that delivers 1W, consumes 2W, but needs only 1mW of input drive (PAE=49.95%). Since most PAs on the market are discrete components, the input drive is a critical concern for designers who plan to drop these discrete components into their designs. At RF, most discrete components are input-matched and output-matched to 50Ω, potentially requiring a large amount of power to drive the input of one of these components. Conversely, if a component is being designed in a single-chip, integrated environment, the input to the component can be made nominally capacitive (especially in CMOS, where the gate input is capacitive), and very little "real" power needs to be consumed to drive it.

Another key issue in the design of PAs is the issue of linearity. With efficiency being so dominant a concern in the end, designers look to boost the efficiency by any means necessary, even if it comes at the expense of another design parameter, like the linearity. In many systems, however, that is not an unreasonable trade-off; the reason being that many communications systems in use today utilize modulation schemes that allow for reduced linearity performance in PAs. In general, modulation schemes can be separated into two basic categories: constant-envelope and non-constant-envelope. In the former, there is no symbol information contained in the amplitude of the transmitted signal, and thus there does not need to be a linear relationship between the input and the output; the output signal will contain only one magnitude at a given time. In the non-constant-envelope case, there is symbol information contained in the transmitted signal, so the PA must accurately amplify the amplitude of the signal that it is driven with. Non-linearity in the PA can cause the transmitted signal to contain the incorrect information, corrupting the communications link.

In the constant-envelope modulation scheme, the symbol information is transmitted in the phase of the transmitted signal. Therefore, what is important is that the signal path not distort the phase of the signal. Unlike the non-constant-envelope case, the signal to be transmitted will have a constant amplitude, since no information is contained therein, and the PA will convert that to a constant amplitude output signal. Even if the PA does not linearly amplify differing input amplitudes, that is acceptable; only one amplitude will be present at any given time. In general, PAs for cellular systems must be able to control their output power level, meaning that they will have to supply different levels of power at different times, but as long as the input-output relationship is known, the input level can be varied to reach the desired output level (which will still be constant amplitude). The power level variations in the mobile unit usually only vary in between frames (transmit times) for constant-envelope modulation schemes; in a given frame, the power level is constant.

Phase distortion in cellular PAs is a concern among designers. In cellular PAs, especially those in the mobile units, phase distortion manifests itself as spectral regrowth [11], which is often characterized by the Adjacent Channel Power Rejection (ACPR) of the PA. In the frequency domain, the power of the desired signal, which is mostly contained in a specified band, will be spread outside that specified band if the phase information is distorted. Unfortunately, most cellular standards have stringent requirements on the amount of power that can be transmitted in adjacent channels, in order to prevent interference between mobile units. It can be shown, however, that the cause of phase modulation in PAs is AM-PM distortion, or Amplitude Modulation-Phase Modulation distortion. Nonlinear PAs will convert possible amplitude modulation in the signals at the input into phase modulation in the output, which will cause phase distortion[11], but if the input is constant amplitude, the phase information will be converted without distortion. For different amplitudes, the phase delay may be different, but because the amplitude is constant in a given frame, the relative phase between symbols will be unchanged, and the signal will be correctly transmitted. As a result; PAs with heavy nonlinearities can be investigated for use in systems using constant-envelope modulation schemes.

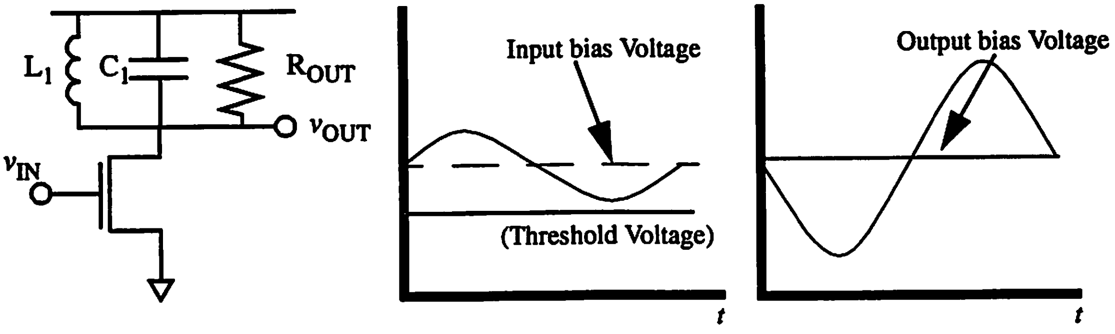
2.4 PA Classes

In general, PAs can be placed into two different categories: one in which the device nominally acts as a current source (i.e. in its amplifying mode), and one in which the device acts as a switch. One common convention is to refer to the former group as the "linear" class of PAs, even though the specific implementation may have a very nonlinear relationship between input and output. The second category is then usually referred to as the "nonlinear" or "switch-mode" class mode of PAs. Each of the categories has several different sub-classes, which are used to identify the topology used in a particular implementation. Other works go into great detail regarding the different PA classes[8][9][10][13], so that work is not repeated here. A brief look at the different categories and sub-classes will be presented here for the sake of background, but the references should be used for a deeper understanding.

2.4.1 "Linear" or Amplification-mode PAs

The term "linear" is enclosed in quotes because the actual input-output relationship of PAs in this category does not have to be linear; as stated earlier, this just identifies the group of PAs in which the device is intended to operate in its amplifying region. For FET devices, that would be the saturation region, whereas for bipolar devices, that would be the forward-active region. Since the devices are meant to operate in their amplifying region, it should be apparent that there will be some relationship between the magnitude of the input and output, regardless of how linear that is.

The big issue in the design of "linear" PAs is the trade-off between linearity (i.e. the linearity of the input-output relationship) and efficiency. The most linear PAs are those in which the amplification device is always conducting current; the instantaneous current through the device is a function of the instantaneous input, to first order. However, in order for this to be true, the DC bias of the input signal must be quite high, to allow for the device to still conduct current when the input swing is at its lowest point. As a result, the average power that is consumed, even when no signal is applied, can be quite high, and the efficiency will suffer. The above description describes the group of PAs known as Class A PAs, in which the amplifying device conducts current for the entire input sinusoid cycle. Figure 2-4. shows the basic implementation and operating conditions of a Class A PA. The amplifying device (shown here as an FET device) is biased in such a way that it always remains in its amplification region, even under maximum input signal conditions. The input bias voltage is set such that the maximum input swing keeps the input signal above the threshold voltage required to keep the device on (Figure 2-4.b). The output voltage swings around its bias point; in general this is the voltage supply, V_{DD}. In an ideal view, the maximum amplitude of the output swing is just VDD, which can help us determine the peak efficiency of the Class A configuration. If we consider that



(a). A Basic Class A Implementation

(b). Input Signal for a Class A PA

(c). Output Signal for a Class A PA

Figure 2-4. Class A PA Operation

$$\hat{I}_o = \textit{average current} \qquad \text{Eq. 2-5}$$

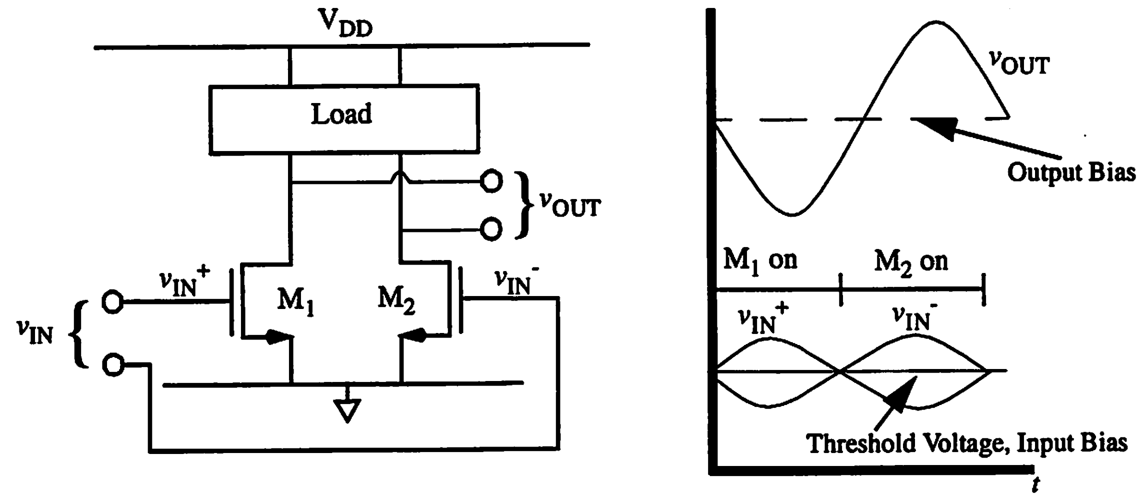
then the peak efficiency is given by

$$\eta = \frac{\frac{1}{2}V_{DD}\hat{I}_o}{V_{DD}\hat{I}_o} = \frac{1}{2} = 50\% \qquad \text{Eq. 2-6}$$

So the peak efficiency of the Class A PA is 50%. However, this is in general not approachable in a realistic implementation. First, most components (both passive and active) will have some amount of real resistive loss in them, and thus there will be some measure of resistive loss due to the current flowing through these parasitic resistances. Moreover, in practice, getting a peak voltage swing of V_{DD} is usually not possible, as the amplification device will leave its amplification region and enter its resistive region (linear for FETs, saturation for bipolar devices). As a result, the peak voltage swing is reduced; these and other real issues limit the attainable efficiencies of fully Class A PA implementations to about 30%.

The next possibility would be to consider a PA with a device that was not conducting current all the time. This is the idea behind the Class B PA, sometimes also referred to as a "push-pull" output stage. In standard implementations, two amplification devices are used, each of which amplifies the signal for half the sinusoidal period. The easiest way to accomplish this is to bias the devices such that they are on the edge of conduction in the quiescent state, and then to have the voltage excursions in each direction turn the appropriate device on. That is, as the voltage starts to go above its steady-state value, one of the devices will turn on; as the voltage falls below its steady-state value, that first device turns off and the second device turns on. Figure 2-5. depicts a common implementation and the standard waveforms of the Class B PA. Because each of the devices is only conducting for one half of the input cycle (and not consuming power for the other half of the cycle), the efficiency of this implementation is greater than that of the Class A implementation. The theoretical peak efficiency of the Class B PA can be calculated to be

$$\eta = \frac{\pi}{4} = 0.78 = 78\% \text{ [12].} \qquad \text{Eq. 2-7}$$



(a) Amplifying, Differential Class B PA

(b). Input Signal for a Class B PA

Figure 2-5. Class B PA Implementation

In practice, efficiencies of the Class B implementation can reach 50% or slightly more. However, the linearity of the PA is degraded in this implementation. Not only are there issues of matching between the two devices (if the gain through the devices is not exactly the same, the output will not be a smooth sinusoid), but if the two devices are not biased exactly at their threshold voltages, the issue of crossover distortion arises. Crossover distortion occurs when both devices are in their off-states when the input signal crosses zero. So while the efficiency of the stage has increased, the linearity has decreased.

One method of achieving somewhat better linearity is to compromise between the Class A and Class B implementations. This is known as the Class AB implementation. The PA is now biased such that it is on for more than half the cycle (but not the entire cycle). In this case, the issue of a "dead" period when the crossover point is reached is avoided, because there is a portion when both devices in a push-pull implementation are on. In narrowband RF implementations, Class B and AB PAs can also be implemented using a single device, as an RF filter at the output can be used to extract the fundamental frequency component of the output waveform. Normally, a single device implementation would pose a problem, in that the device would be off for part of the input cycle, giving a chopped, and thus extremely distorted, waveform. However, through the use of narrowband RF filters, the component of the output waveform at the fundamental frequency can be extracted, and the amount of distortion can be reduced.

It is precisely because of this narrowband nature that it is possible to use a PA in which the bias voltage is further reduced, *below* the turn-on voltage of the device. This is the Class C PA, the focus of this work. The Class C PA will be discussed in great detail in Section 2.5, but briefly, the PA is biased below its turn-on voltage, and the input drives the device on for a small portion (less than half) of the input cycle. This creates a pulsed current in the device; this pulsed current is then filtered, so as to extract the fundamental frequency component, which is then passed to the resistive load, creating an output waveform that is at the fundamental frequency. Again, this will be discussed in more detail in Section 2.5.

2.4.2 "Nonlinear" or Switch-mode PAs

The group of Nonlinear PAs is also known by a more descriptive name: switch-mode PAs. These are the PAs in which the device is meant to act as a switch. For RF PAs, the two classes of switched mode PAs which have received the most attention are Class D PAs and Class E PAs. If the process of signal amplification is thought of as a power conversion process, then switched mode techniques used in power conversion systems, such as DC-DC converters or regulators, can be used in these PA applications. That is the heart of the Class D and Class E architectures.

The Class D PA is the first of these switched-mode classes of PAs, and has recently been implemented in a CMOS implementation[15]. In the Class D PA, current from the supply is steered between the device, when the switch is closed, and the load, when the switch is open. The Class D architecture is similar to what is used in a bridge

DC-DC converter[13]. In that style of DC-DC converter, devices acting as switches change the polarity of the input voltage onto the load, and the resulting output is averaged to create a output voltage that is some fraction of the input voltage, depending on the duty cycle of the switching. This push-pull action can also be used in the design of an RF PA, where the load is connected to switches which switch the voltage across the load. If the switching is done at the output carrier frequency, the narrowband nature of the transmitted signal allows the use of the RF filters to pass only the fundamental frequency component. Again, because of the ability to filter out unwanted components of the output signal, this type of amplification can be done with only one device, in which case the power from the supply is either sunk in the device or the load. Because of the use of a series L-C circuit tuned to the output frequency, the current in the device will be a sinusoid for the period that it conducts current. If two devices are used, each will carry a half-sinusoidal current waveform (each will be on for half of each cycle). If the implementation of the switch is assumed to be ideal, i.e. no on-resistance and the output voltage is exactly zero when the switch is closed, the ideal maximum efficiency of the power stage can be 100%, as no power will be consumed in the transistor. However, due to non-zero on-resistance, the maximum attainable efficiencies can be dramatically reduced, especially in CMOS implementations; one published result of a Class D PA in CMOS shows a drain efficiency of 62%[15].

The Class E PA, while like the Class D PA, uses the idea of soft switching in order to further reduce any power consumed by the device in the switched-mode PA. This class of PA has also been recently implemented in a CMOS implementation[16]. Essentially, the Class E PA tries to force the voltage on the output node to zero at the instant that the switch is closed, so that there is ideally no time at that transition when both the output voltage and current are non-zero[14]. Not only that, but there is also no CV^2 energy loss from the output capacitance discharging as the switch is turned on. Similarly, the Class E PA is designed to force the current flowing in the switch to be zero at the instant that the switch opens. Again, this is to ensure that there is no period of time around that transition when both the current and voltage are non-zero. Also, in order to account for timing errors in the switching instants, the slope of the output voltage waveform should also be zero at the instant of that the switch closes. Thus if there is any timing error in when the switch closes, the power consumed because of any overlap in the output current and voltage waveforms will be minimal; since the slope of the output voltage at the correct instant is zero, the value of the output voltage at instants close to the output voltage will be very small. The basic operation and voltage waveforms of the Class E PA can be seen in Figure 2-6[16]. In the figure, the switch is used to represent the transistor. Like the Class D PA, the theoretical peak efficiency of the Class E PA is 100%; again, practical considerations, especially in CMOS limitations, have limited the efficiency to about 50%[16], although GaAs implementations have reached close to 60% efficiency[17].

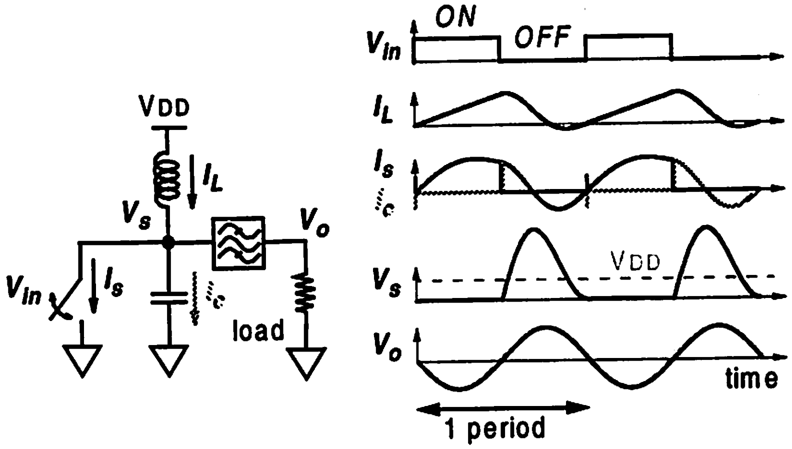


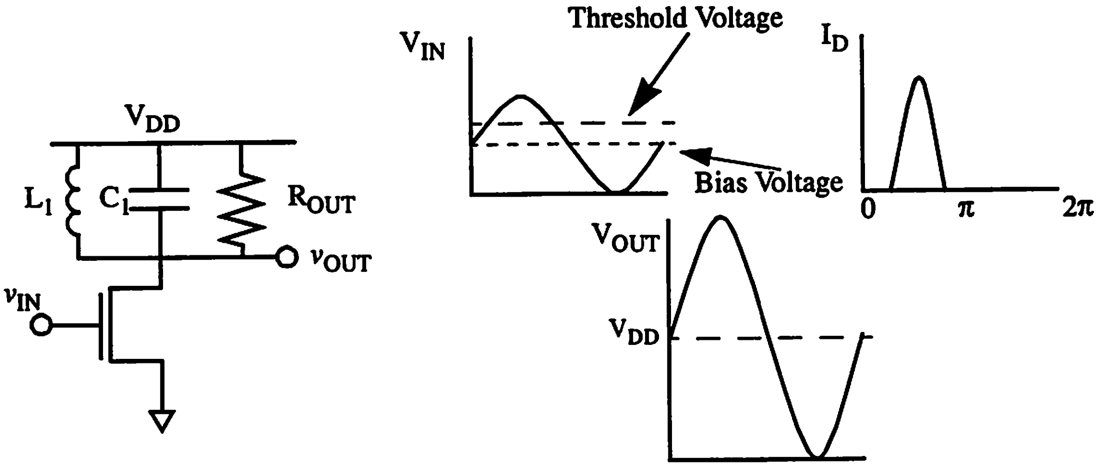
Figure 2-6. Class E PA Implementation and Waveforms

2.5 Class C Power Amplifiers

The focus of this work is the design and implementation of Class C Power Amplifiers; in this section, the operation of Class C PAs will be discussed in detail. As state earlier, the Class C PA takes the progression from Class A to Class B PAs and carries it even further. The simple explanation is to say that the device is biased below its threshold or turn-on voltage; the input drive turns the device on for a small portion of the input sinusoidal cycle. However, in order to completely understand what happens in the Class C PA, a deeper investigation must be undertaken; this investigation must not only include the Class C mode of operation as well as what happens when a specific type of device is operated under those conditions.

2.5.1 Idealized Analysis

The basic implementation and operation of the Class C PA is shown in Figure 2-7. The PA is biased below its threshold voltage, and the input signal drive turns the device for a fraction of the input signal cycle. This in turn generates a current that is pulsed in the device; the current is then filtered, with the fundamental frequency component going through the desired load. Unlike a switched-mode PA, the theoretical maximum efficiency is not 100%, which can be seen from the fact that the current and output voltage waveforms are both non-zero at the same time; the device does consume power. However, it can be seen that this overlap region is considerably smaller than that in the Class A or B cases, and the voltage is close to zero while the current is non-



(a) Single-Ended Class C PA (b) Class C PA Waveforms

Figure 2-7. Class C PA Implementation and Waveforms

zero, if the output voltage is indeed swinging almost down to ground. As a result, the efficiency can be considerably higher than in the other "linear" cases.

To further understand the operation of the Class C PA, a simple analysis can be done, using an idealized model. First assume that the device to be used has some function that relates the output current and the input voltage, as in Eq. 2-8:

$$I_{DEVICE} = K f(v_{IN}) . \qquad \text{Eq. 2-8}$$

Assuming further that the input is sinusoidal, the current over one period can be defined by the following piecewise equation:

$$I_{DEVICE} = \begin{matrix} 0 & \textit{for } v_{IN} < \textit{turn-on voltage} \\ K f(v_{IN}) & \textit{for } v_{IN} > \textit{turn-on voltage} \end{matrix} . \qquad \text{Eq. 2-9}$$

In the narrowband RF case, the assumption is made that a narrowband RF filter can be used to extract the fundamental frequency component of the current while suppressing the higher order harmonics. Thus, using a Fourier Series expansion, the current delivered to the load is given by

$$\hat{I}_{LOAD} = \frac{1}{T}\int(\sin\tau)(I_{DEVICE}) = \frac{1}{2\pi}\int_{\tau_1}^{\tau_2}(\sin\tau)(K f(v_{IN}))d\tau \ , \qquad \text{Eq. 2-10}$$

where τ_2 and τ_1 are the endpoints of the period during the cycle when the voltage is above the turn-on voltage and thus current flows through the device.

Ideally, the load is a resistive one, so the output voltage is known to be

$$v_{OUT} = \hat{I}_{LOAD}R_{LOAD}\sin\tau . \qquad \text{Eq. 2-11}$$

Knowing this, the output power and efficiency can be calculated. The output power is simply

$$P_{OUT} = \frac{\hat{I}_{LOAD}^{\ 2}}{2}R_{LOAD} \qquad \text{Eq. 2-12}$$

Finally, in order to calculate the efficiency, the total power drawn from the supply must be calculated. The active device is the only element of the circuit that draws current from the supply, so the total power consumed from the supply is

$$P_{DC} = V_{DD}I_{AVG} = \frac{V_{DD}}{2\pi}\int_{\omega_1}^{\omega_2} K f(v_{IN}) d\omega . \qquad \text{Eq. 2-13}$$

As stated earlier, the efficiency is just the ratio of P_{OUT} to P_{DC}. However, as will be seen in the next section, this is purely an ideal and theoretical analysis. Device non-idealities make the full analysis much more difficult.

2.5.2 Class C PA Non-Idealities

Stated simply, there are many nonidealities in the Class C PA that are present when a real device is used in the implementation. Among the factors that must be accounted for are the finite gain in the device, region of operation of the device, and the nonlinear output capacitance of the device. Each of these factors will be discussed in detail in this section.

First, the finite gain of any real device which is used in this class of PAs becomes a problem because the amount of current that is generated in the device is small. An examination of the output current waveform makes this easy to understand.

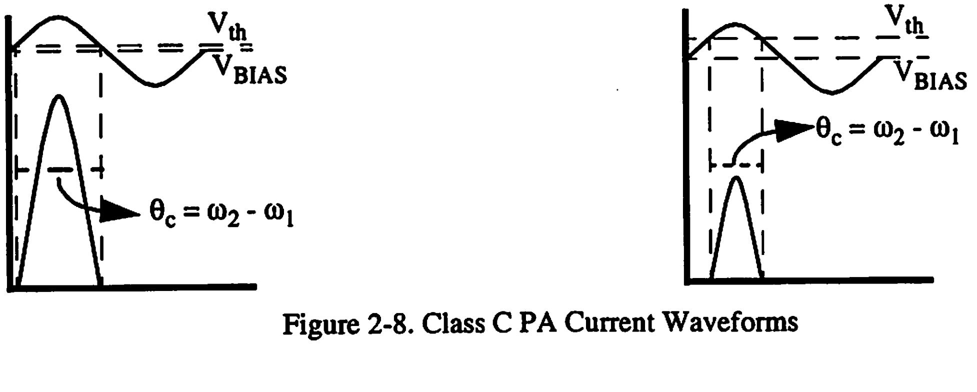

V_{th}
V_{BIAS}
$\theta_c = \omega_2 - \omega_1$
V_{th}
V_{BIAS}
$\theta_c = \omega_2 - \omega_1$

Figure 2-8. Class C PA Current Waveforms

Sample waveforms are shown in Figure 2-8. θ_c is defined as the conduction angle, which defines the portion of the cycle that the device is on, or conducting; it is equivalent to the difference between τ_2 and τ_1 used in the previous section. As the conduction angle is reduced, the amount of current through the device is reduced as well. More importantly, the magnitude of the fundamental frequency component, which determines the power delivered to the load, is reduced as well. However, the efficiency is increased at the same time. Taken to its logical extension, this says that as θ_c approaches zero, the efficiency tends to 100%, but the power delivered to the load tends to zero as well! Since this extreme case is more of a quirk than anything else, a realistic Class C PA will never get 100% efficiency and any semblance of output power; essentially, infinite gain would be needed to get non-zero output power in that case.

The second issue that must be considered is the region of operation of the device, which becomes especially critical in the case of a MOS implementation. To a first approximation, the region of operation of a MOS device is determined by its terminal voltages. Assuming that the source and bulk are at the same potential, the device conducts current when its gate-source voltage, V_{GS}, is greater than its threshold voltage V_T. It is in the saturation region (in which it is best used as an amplifier) when its V_{GS} is greater than V_T and its drain source voltage, V_{DS}, is greater than the difference between the V_{GS} and V_T. When V_{DS} is less than the difference between the V_{GS} and V_T, the device is in the triode or linear region. Ideally, an MOS PA which is supposed to be a "linear" PA should have its device in the saturation region while it is on. However, under conditions where the output swing is at its maximum, the terminal voltages may violate that rule, as shown in Figure 2-9. Under conditions of maximum swing, the output voltage, and thus the V_{DS} of the device, will be less than the

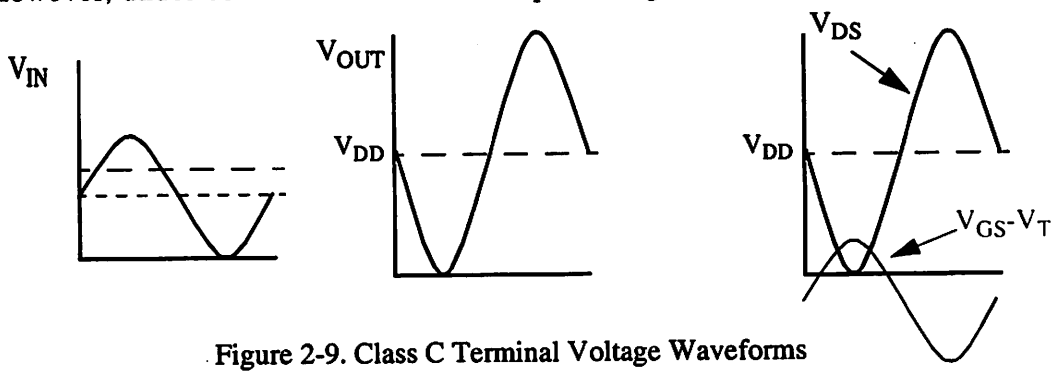


Figure 2-9. Class C Terminal Voltage Waveforms

difference of the V_{GS} and the V_T of the device, forcing the device into the triode region for a portion of the input cycle. As a result, the operation of the entire circuit is affected during this peak condition, and cannot be so easily described. This mode of Class-C operation is known as mixed-mode Class C operation[10], as opposed to "true" Class C operation, or the case described in the idealized section (the device is always in its amplification mode when on).

2.5.3 Benefits of Class C PAs

In the previous sections, it has been established that Class C PAs can not, even theoretically, reach 100% efficiency with any real power output, and that at maximum power output conditions, the Class C PA so deviates from ideal operation that it is very difficult to accurately describe the operation. The question that arises next is why such a PA should be investigated then; switched-mode PAs may offer equal or better efficiencies, without the design complexities and uncertainties that come with the Class C PA. The answer to this lies in the relationship between input amplitude and output amplitude of the respective PAs. In the Class C case, while the input-output relationship is nonlinear, the output amplitude still varies with a varied input level. However, in the case of switched-mode PAs, the output amplitude is fixed relative to the input amplitude. Since the device used in the PA is meant to act like a switch, the input signal, once large enough to close the switch, does not further affect the output amplitude, to first order. Because switched-mode PAs work by "switching" the voltage across the load between the supply and ground, the only way to control the level of the output is by controlling the supply voltage; therefore, in switched-mode PAs, an agile voltage regulator is needed in order to control the output power level. That is one more block that needs to be designed, and one more block that directly reduces the efficiency of the PA. While constant-envelope modulation schemes have constant power while transmitting in a given frame, most cellular communications schemes require that the mobile PA be able to transmit different power levels at different times, depending on the conditions around the mobile user. The Class C PA can do this simply by varying its input signal level. This advantage, added to the efficiency improvement that can be gained over standard Class A or B PAs, makes this investigation important.

Even more importantly, the future of communications systems is moving towards non-constant envelope modulation schemes. As the demand for higher data-rates in wireless systems becomes overwhelming, the next generation systems are moving towards including data in both the amplitude as well as phase of the transmitted signals, requiring very linear PAs. There are two ways to improve PA linearity. The first is to use an extremely linear, but highly inefficient PA, like a Class A PA. This solves the linearity problem, but dramatically reduces the battery life of a portable unit. The second solution is to continue to use a higher efficiency PA architecture (like Class C or higher), but then to try to improve the linearity of the PA or the entire transmitter. There are several schemes for transmitter linearization, including Cartesian Feedback [18], Digital Predistortion[19][20], and Feed Forward[21], among others. One benefit of the Class C PA over the switched-mode PAs in this case is that because of the input-output amplitude relationship, Class C PAs can be used with methods like Cartesian Feedback, which vary the amplitude of the input.

Therefore, an investigation into the design of a Class C PA is needed; the Class C PA does have benefits over the more linear types of PAs (i.e. better efficiency), as well as over the switched-mode classes of PAs, as stated above.

2.5.4 What Is Needed

In order to design and implement a Class C PA, a lot of work is needed. First and foremost, the issue of the mixed-mode PA must be addressed. To this point, no clear method exists to design a mixed-mode Class C PA. It would be beneficial if some type of design methodology or design guidelines existed, whereby a designer could avoid the blind utilization of circuit analysis software as the sole means of reaching the desired design goal. Currently, the method used to design extremely non-linear amplification-mode PAs is extremely empirical. A simple flow chart indicating the standard procedure is shown in Figure 2-10. An optimum load resistance is determined

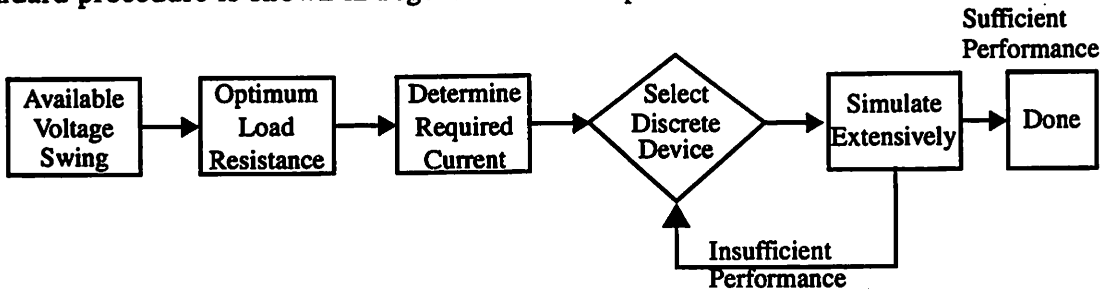


Figure 2-10. Standard Non-linear PA Design Procedure

from the voltage swing available of the output of the device. In general, these PAs are designed using discrete transistor devices; as a result, each of these devices is extremely well-categorized and its performance is well-known. A device is selected, and a corresponding output matching network is selected as well. The PA is simulated; if the performance meets the desired requirements, the design is done. If not, a different device is chosen (or different design parameters, such as bias, etc.) and the process is started again.

In the case of an integrated CMOS implementation, however, this is not really a viable solution to the problem of PA design. The type of characterization available for the discrete devices mentioned above is not available for integrated devices. Furthermore, while in the case of discrete devices the set of devices to choose from is finite, in the case of a CMOS PA, a whole continuum of possible devices is available, as the device size (both width and length) is an entirely free variable in the design. As a result, a less exact, scalable model is the only one that is available. Using the method described above in this process would be extremely difficult, as there is no a priori method of limiting the device choices as there is in the discrete case. As a result, another method more suited to the CMOS case must be investigated. While the large-signal operation of the circuit may be too complex to come up with a single simple, closed-form equation that encapsulates the entire problem, a simple method that gives a designer an understanding of the trade-offs between certain design parameters and allows the designer to significantly narrow his focus before reaching the stage of using the circuit analysis software would be very beneficial.

The goal of the next chapter, therefore, is to present just that; a simple design methodology which gives the designer a quick way to reach a reasonable solution which can be used as a starting point in a circuit simulator. Instead of having to use a circuit simulator to understand all the design trade-offs inherent in varying certain design parameters, the design methodology will help the designer do that, and leave the resource-intensive circuit simulator for "tweaking" the design around a design point reached through the methodology.

3

Class C Theoretical Analysis

3.1 Introduction

In Chapter 2, various classes of PAs were introduced, with a focus on the Class C PA, which is the focus of this work. As stated in Chapter 2, a design methodology for designing Class C PAs, especially those in which the device might leave its amplification region and enter its "switch-like" region of operation, isn't currently available. In this chapter, a simplified design methodology will be described. In order to remove some of the complexity inherent in the methodology, this work will focus on using MOSFETs as the active devices in the PA. The method for designing the PA will be built on a straight-forward model of the MOS device, as explained in the next section. Reasons why that particular model may be used in this work will also be explained. Once the MOSFET model has been introduced, the steps involved in developing a valid PA design that meets all the required specifications will then be introduced. First, the general method will be examined, after which the method will be applied to a sample design. Furthermore, the work in this section will initially focus on the design of the "output" stage of the PA; that is, the stage which actually drives the output (in most cases, the RF filter and the antenna). In general, the design of the earlier stages of the PA is less complicated than that of the output stage. Methods for designing those stages will be addressed at the end of this chapter.

3.2 CMOS Device Model

3.2.1 CMOS Device Basics

The MOS device is a *field-effect transistor*, i.e., the electric fields in the device are responsible for the current flow. A cross-section of a standard device is shown in Figure 3-1(a). The device has four distinct electrical terminals: the drain (D), the gate (G), the source (S), and the bulk (B); the electrical symbol for the transistor is shown in Figure 3-1(b). Current flows from the drain to the source, and is controlled primarily by the signal applied to the gate, but also to a lesser extent by the signal on the bulk. The simplest model for this flow of current is the first-order *square-law* model defining the relationship between the signals at the 4 terminals. This model in its simplest form states that the current flow between the drain and the source, I_{DS}, is related to the gate-to-source voltage V_{GS} and drain-to-source voltage V_{DS} in the following manner:

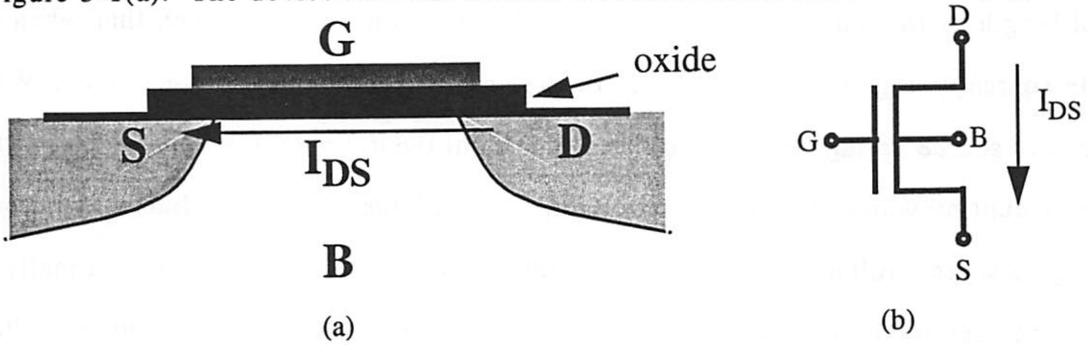


Figure 3-1. MOS device cross-section

$$I_{DS} = \begin{cases} 0 & V_{GS} < V_T \\ \mu_n C_{OX} \frac{W}{L}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right) V_{DS} & V_{GS} > V_T \, and \, V_{DS} < V_{GS} - V_T . \\ \frac{\mu_n C_{OX} W}{2 \quad L}(V_{GS} - V_T)^2 (1 + \lambda V_{DS}) & else \end{cases} \qquad \text{Eq. 3-1}$$

V_T is the threshold voltage for the MOS device, which is determined by the material properties and geometries of the device. μ_n is the electron mobility, which is a constant describing the ability of electrons (or holes, if holes were the free carrier in the device of interest) to travel in the bulk material (in this work, silicon; electrons have a significantly higher mobility in Gallium Arsenide). C_{OX} is the capacitance inherent in the gate oxide, which is just the dielectric constant of the oxide (ε_{OX}) divided by the thickness of that oxide (t_{OX}) in the particular device of interest. W and L are the width and length of the particular device. From the equation, it can be seen that when the gate-source voltage is below a certain threshold, no current flows in the device. When the gate-source voltage exceeds that threshold, but the drain-source voltage is less than the amount by which the gate source voltage is "on," the current is a linear function of the gate-source voltage (and also dependent on the drain-source voltage). Finally, in the final region of operation, the current is dependent on the gate-source voltage *squared*, and also approximately independent of the drain-source voltage, especially if the ($1+\lambda V_{DS}$) term is ignored. While this can be an important component in an exact analysis, as an approximation, it is useful to ignore that term in this work, as will be shown later in this chapter. These three regions are known as cutoff, linear or triode, and saturation, respectively. The saturation region is the one in which the gate voltage is most amplified in the form of the output current. Ideally, for any application in which an incoming signal is to be amplified, the MOS transistor should always remain in the saturation region, thereby affecting the largest gain.

However, as the amount of current that the transistor drives increases, so does the magnitude of the signal on the drain of the transistor, potentially causing the drain voltage to be pulled below the gate-source voltage. At this instant, the device moves into the linear region, and now the current is also dependent on the drain voltage (and not primarily controlled by the gate-source voltage). As an example, consider the simple single transistor circuit shown in Figure 3-2(a). As the magnitude of the input sinusoid increases, the magnitude of the current flowing from drain to source does as well, and at some point, the output voltage will become so large that the transistor will be forced into the linear region. What exacerbates this problem is that the MOS device is an *inverting* device; that is, the output signal has the opposite polarity of the input device. The problem itself is illustrated in Figure 3-2(b): as the input sinusoid reaches its maximum, the output sinusoid reaches its minimum. This forces the device into the linear region much earlier than if, say, the device was a non-inverting device and the input and output voltage waveforms were in phase.

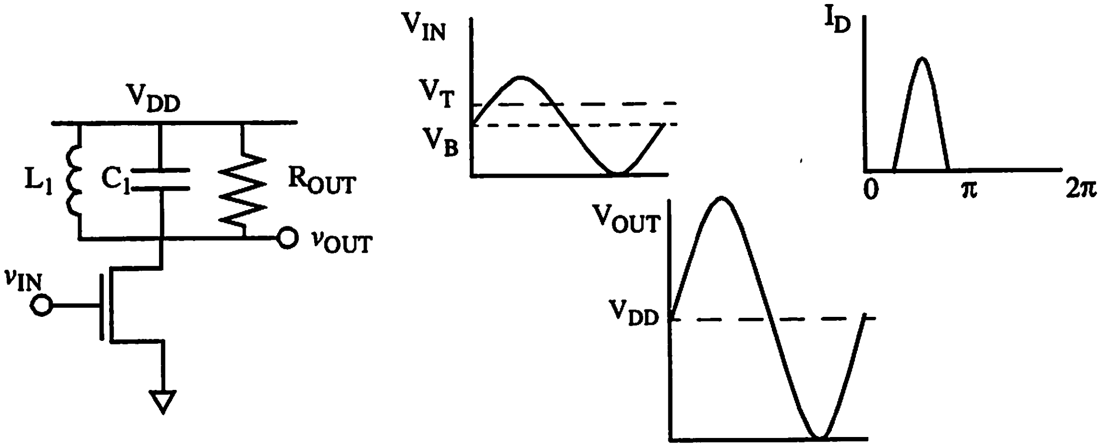


Figure 3-2. Single Transistor Circuit and Waveforms

All of this is important because it does not allow the circuit to be analyzed in a simple linear method in which the input and output are decoupled. In many circuit designs, the design process begins with a linearized model of the transistor to be used, incorporating well-known small signal parameters such as g_m and r_o. These linear elements can then be analyzed to come up with closed-form relationships for different circuit topologies, allowing for the design process to go forward and for a designer to approach the circuit solution for a set of specifications. However, in the case of a power amplifier (PA), the design is significantly different. First, the small-signal linearized model of the transistor cannot be used in the design process; the signals in the output stage of the PA are too large. Second, the transistor used in the PA can potentially operate in all 3 of its distinct regions of operation. If it could be guaranteed that the transistor would remain in the saturation region while on (the "true" class C operation described in Section 2.5), a closed form solution might be approximated, as the relevant equations for the drain current do not depend on the drain voltage. However, the transistor will be forced into the triode region as well, preventing a closed-form solution from being reached. Another way to understand the difficulty is to realize that in the triode region, the MOS transistor is thought of as a variable resistor. In other words, the transistor transforms from an open circuit (turn-off region) to an amplifier (saturation region) to a resistor (triode region). Any design methodology described in this section must account for these changes.

3.2.2 Validity of Model

Certainly, the model described in Section 3.2.1 is an extremely simplified one; the current models that are used in circuit analysis programs like HSPICE, such as BSIM3, are significantly more complex, and are usually based on data taken from actual devices during a device characterization phase. A transistor model like BSIM3 can have as many as 25 parameters[31], making it somewhat unwieldy for use in a simple design process. Because the goal here is to come up with a procedure that allows the designer to generate a workable design without first resorting to circuit analysis tools like SPICE, the more advanced models should be removed from consideration. In this chapter, two additional simplifications will be made to the square law model presented in Eq. 3-1: first, the $(1+\lambda V_{DS})$ term in the saturation-region equation for the drain current will be dropped, in order that the input voltage and output voltage will be decoupled while the device is in saturation, and second, an average value will be used for the variable drain-bulk capacitance, which is a common approximation made in these designs in any case. The first approximation is not as problematic as it might seem, as in this particular case, the device is primarily conducting current while the output voltage is low, so the V_{DS} term will generally be small and decreasing during the time that the device is on, reducing the error due to this approximation.

It is also important to investigate where the simple square-law model deviates from reality, and to understand if and how this deviation affects the design process or methodology that will be investigated in this chapter. In a short-channel MOS device, the operation of the device can deviate greatly from the predictions of the square-law model, especially under certain circumstances where the device is under great strains due to the signals applied at the device terminals. A few of the potential deviations from the basic square-law model will be examined in the upcoming sections.

3.2.2.1 Short Channel Effects on Threshold Voltage

One of the critical parameters of the device model that is affected by the continuing reduction in device geometries is the threshold voltage, V_T[22]. As the device channel-length decreases, the depletion regions surrounding the drain-bulk and source-bulk regions take up more and more of the channel. As a result, the amount of charge needed to cause the inset of inversion is reduced, and thus the gate voltage required to create a channel is also reduced. This can be modeled as a reduction in the threshold voltage required to turn the device on. Furthermore, as the drain voltage is increased (increasing the reverse-bias of the drain-bulk junction), the size of the depletion region in the channel increases, further reducing the gate voltage necessary to invert the channel and cause current flow. This effect is known as *Drain-Induced Barrier Lowering* (DIBL)[32].

However, in the case of the particular architecture in this work, a critical point to be noted is that while the device is conducting current, the drain voltage will generally be quite small. Since the MOS device is an inverting device, while the gate voltage is at its maximum, the drain voltage is at its minimum. Therefore, the effect of DIBL is ameliorated. As long as the peak voltage at the drain (which occurs when the device is off) can be supported without damaging the device, this effect should not impact the standard square-law model described earlier.

3.2.2.2 Short Channel Effects on Drain Current

The short channel nature of sub-micron CMOS can also affect the current flowing in the channel, through its effect on two critical parameters. First is the effect on the mobility of the free carriers in the channel (μ), and the second is the effect on the velocity of the carriers in the channel. Since this work is primarily concerned with N-type devices, in which electrons are the free carriers, the terms free electrons and free carriers will be used to mean the same thing.

As the electric fields in the channel increase, both the mobility and velocity of the free electrons is adversely affected. In the case of the mobility, the transverse electric field (caused by the gate voltage) is the source of the degradation. At low transverse electric fields, the mobility is the factor which relates the strength of the electric field and the free carrier drift velocity. However, as the transverse electric field increases, the free electrons undergo an increased number of collisions with the surface of the silicon (the interface between the silicon and the oxide). This degrades the mobility of the carriers at the surface relative to their mobility in the bulk of the silicon.

It is important to remember that it is the gate voltage overdrive that causes this degradation in the mobility; one way to limit the effect of this mobility degradation is to ensure that the gate overdrive is not overly excessive. That is, if a restriction that the gate overdrive should not reach a certain level is adhered to, the simple square-law model may still be valid for the first order analysis to be undertaken. Interestingly enough, this is a limit that is actually encouraged by the technology! As described in

Chapter 4, CMOS is inherently a poor amplification technology, and as such, obtaining an extremely large overdrive signal can be difficult and will require a large power consumption in the pre-amplification stages. In order to maintain a high efficiency, a more moderate drive level is beneficial, and thus the transverse fields in the channel should not become excessive. This will mitigate the effect of gate-induced mobility degradation, making the simple square-law model more applicable.

The second route through which the drain current is affected is through the mechanism of velocity saturation. As the lateral electric field, due to the drain voltage, is increased, the velocity of the carriers saturates (at about 10^7 cm/s), reducing the drain current below what might be expected from the classic square-law model presented earlier. However, as stated earlier, because the drain voltage will be low when the device is conducting current, to first order, the issue of velocity saturation should not be too detrimental, and as such, the square-law approximation should not be too inaccurate. Since the purpose of this analysis is not to provide a 100% accurate construction of the operation of the PA but rather a first order estimate of the PA behavior and key performance metrics, using a simplified model which is accurate to first-order but somewhat less accurate under heavier scrutiny is still valid.

3.3 Class C Circuit Basics

Section 2.5 analyzed the operation of a "true" Class C Power Amplifier (PA), and further went on to describe the operation of a mixed-mode Class C PA and the difficulties in designing such a PA. In the previous section, a simple device model for the MOS transistor was described and the validity of said model for use in an *a priori* design process was confirmed. So now the question of how to actually begin designing a PA (and thus a method for that design) must be addressed. What first must be investigated is the circuit topology and the limitations that impact the design, which are needed to start the design process.

3.3.1 The Circuit

The first step in the process is to determine the circuit to be analyzed. As described in Section 2.5, the simplest Class C PA is of the architecture shown in Figure 3-3(a). Shown in the figure is only what is generally the final stage of a PA, which consists of a single transistor and a tuned output circuit. The inductor and capacitor at the output are tuned to resonate at the RF carrier frequency, which forces all the current at that frequency into the resistive load.

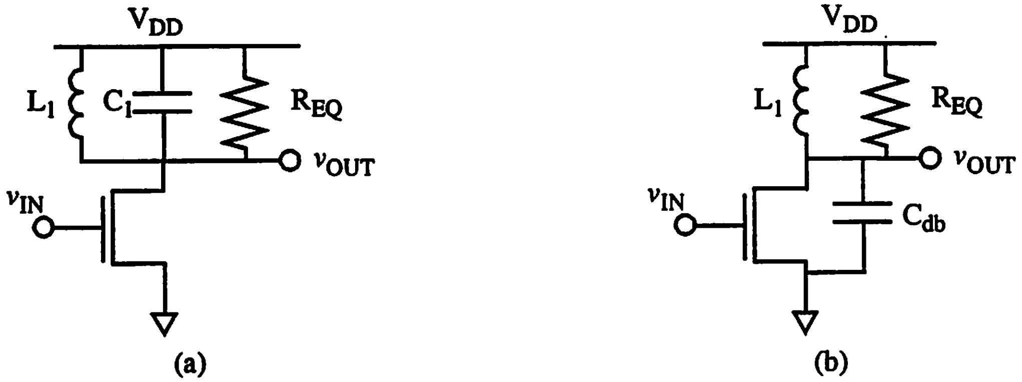


Figure 3-3. Class C PA

However, the transistor in Figure 3-3(a) has parasitic components that must be accounted for when analyzing the circuit. First and foremost, the CMOS transistor has a significant non-linear capacitance between the drain and bulk. Because the bulk connection is generally connected electrically to the source, that capacitance appears between the drain and ground. This capacitor, especially in the case of CMOS, can be so large as to encompass the entire capacitance of the LC tank. As a result, the actual circuit implementation may end up looking like the circuit in Figure 3-3(b). In this case, the inductor actually has a second function; not only does it serve to resonate the drain capacitance, but it also serves to bias the circuit. It provides the DC bias point to the drain, and allows the resonance of the output to occur around the supply, thus offering a greater amount of headroom than if the output were resistively biased.

Furthermore, an idea of the size of the device to be used in this design must be reached. The size of the device will determine part of the output network, as mentioned in the following sections. Without any other restrictions, the device size would be a variable that could be changed arbitrarily to determine the optimum solution. However, this is not the case with a sub-micron CMOS process. As detailed in Chapter 4, the output voltage magnitude is limited by the oxide breakdown voltage of the MOS device, and thus the minimum amount of current needed to generate the required output power is set. The limited voltage swing sets the optimum load resistance to be approximately

$$R_{OPT} \approx \frac{(V_{MAX})^2}{2P_{RF}}, \qquad \text{Eq. 3-2}$$

where P_{RF} is the desired peak output power and V_{MAX} is the available peak voltage swing. This in turn sets the magnitude of the current waveform that must pass through the resistor to be

$$I_{RF} = \sqrt{\frac{2P_{RF}}{R_{OPT}}}. \qquad \text{Eq. 3-3}$$

This will narrow the range of potential values for the size of the device to be used, as the device must be able to generate this much current at the carrier frequency. The device size will be necessary when determining the output network, which will be detailed in Section 3.3.3. The parasitic drain capacitance of the device will determine the magnitude of the parallel inductor used to tune it out. Since the drain capacitance is a direct function of the size of the device, the inductor value required can be determined from the size of the device.

3.3.2 The Input

Now that the actual circuit has been realized, the next step is to understand the exact nature of the input waveforms. As stated in Section 2.5.1, the Class C PA has an input waveform which is above the transistor threshold for less than half the input period, resulting in a pulsed output current. These waveforms are reproduced here in Figure 3-4. However, it should be noted that the waveforms in Figure 3-4 are only

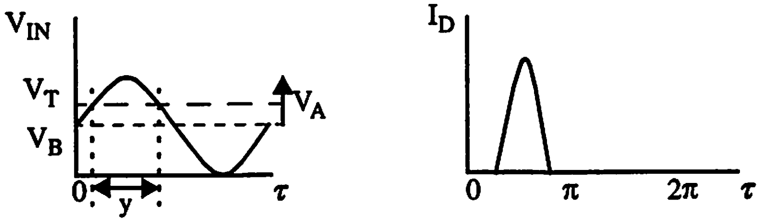


Figure 3-4. Class C PA Waveforms

approximate; the current waveform shown assumes a "true" Class C PA. While the mixed-mode Class C PA will also conduct current during the same times as the "true" Class C PA, the fact that the MOS transistor may enter the triode region for a portion of its on-time will cause the actual current waveform to differ from that shown. The goal of this chapter is to create a method that may predict the current waveform, in order to simplify the design process.

It should be apparent that because of its nonlinear nature, the Class C PA will generate considerable energy at not only the RF carrier frequency (also referred to as the fundamental frequency), but also at harmonics of the carrier. Ideally, the higher-order harmonics would be shunted into the tank circuit and no harmonic components would flow in the resistive. However, because the components that make up the LC tank are decidedly non-ideal, the tank will not completely cancel the other harmonic components at the output. The PA output will be the sum of all of the frequency components at the output as they pass through the non-ideal tank circuit.

One method of analyzing all that is happening in the circuit is to investigate the problem in a purely frequency-domain manner. As is well known, any periodic signal can be broken down and represented as sum of sinusoids of different frequencies. In this manner, the input and output waveforms might be broken down to assist in the realization of the desired design methodology. The Fourier series coefficients for a given signal can be obtained by applying the following equations:

$$b_0 = \frac{1}{2T}\int_T F(\omega)d\omega \qquad \text{Eq. 3-4}$$

$$a_n = \frac{1}{T}\int_T F(\omega)\sin(n\omega)d\omega \qquad \text{Eq. 3-5}$$

$$b_n = \frac{1}{T}\int_T F(\omega)\cos(n\omega)d\omega \qquad \text{Eq. 3-6}$$

where $F(\omega)$ is the function to be broken up into its Fourier components. The function then can be written as

$$F(\omega) = b_0 + \sum_{n=1}^{\infty}(a_n \sin(n\omega) + b_n \cos(n\omega)). \qquad \text{Eq. 3-7}$$

Because the coefficients generally get smaller in magnitude as the value of n increases, it is often enough to calculate the first few coefficients and use those few harmonics to approximate the desired signal.

In the specific case of the Class C PA, the input waveform can first be investigated. While the device is off ($V_{GS}<V_T$), the current flowing in the output is zero, independent of the actual value of the input. When the device is on, the actual amount of overdrive (the device V_{GS}-V_T) is considerably less than the amplitude of the sinusoidal input, since the device is biased below V_T. In other words, the device V_{GS}-V_T can be written as

$$V_{GS} - V_T = V_B + V_A \sin\theta - V_T, \qquad \text{Eq. 3-8}$$

where V_B is the bias voltage (less than V_T) and V_A is the amplitude of the sinusoidal input. Because V_B is less than V_T, the above equation is positive (i.e. the device conducts current) for certain values of θ bounded by

$$\mathrm{asin}\left(\frac{V_T - V_B}{V_A}\right) < \theta < \pi - \mathrm{asin}\left(\frac{V_T - V_B}{V_A}\right). \qquad \text{Eq. 3-9}$$

In other words, the device is turned on for a portion of the cycle centered around π/2. Only the overdrive portion of the voltage causes any current to flow, so when doing the Fourier series expansion, that should be the signal that is considered. Therefore, as an example, the fundamental frequency coefficients of the Fourier Series can be determined by setting n=1 in Eq. 3-5 and Eq. 3-6, as in

$$a_1 = \int_{\theta_1}^{\theta_2} (V_B + V_A sin\theta - V_T) sin\theta d\theta . \qquad \text{Eq. 3-10}$$

In the equation, θ_1 and θ_2 represent the endpoints of the turn-on region of the MOS device, given in Eq. 3-9.

In this manner, the actual overdrive signal can be broken down into its Fourier components. However, analyzing this particular signal is not all that valuable, because the signal then passes through the nonlinear MOS device and produces an output current. If the square-law transfer function were applied to the version of the input overdrive signal which was a sum of sinusoids at different frequencies, the output would consist of a large number of cross-products due to the different frequency sinusoids being multiplied together. A simpler solution might be to apply the square-law transfer function to the input signal and then investigate the Fourier components of the output current waveform, since that is the signal of interest. Furthermore, it is the fundamental component of the output current waveform that ideally passes to the load; the LC tank at the output will theoretically shunt all other harmonics to AC ground.

3.3.3 The Output

The output of the PA has several components, including the output of the device, as well as the components that make up the output network, which are primarily passive components. Each of those components must be understood before the circuit can be analyzed; if the structure of the output is unknown, it is virtually impossible for the operation of the output to be understood.

3.3.3.1 The Output Network

The output network is generally some combination of passive components which are to be used between the transistor output and the antenna. Generally speaking, the antenna can be modeled (to first-order) as a 50Ω resistor. Furthermore, due to the limited voltage swing available at the output, detailed in Section 4.2, that 50Ω load is not the optimum resistance that the device needs to see in order to deliver the required amount of power. As a result, an impedance transformation network which transforms the 50Ω otherwise seen by the device into the optimum resistance needs to be placed between the device and the antenna. A simple output network is shown inFigure 3-5. A step-down network is used to convert the 50Ω load into a lower resistance at the output of the device. A parallel LC tank is used at the output to force the fundamental frequency component of the output current into the "load", which now consists of the antenna and the impedance transformation network. As stated earlier in this chapter, the capacitor in the LC tank is simply the drain-bulk junction capacitor of the device; the inductor simply resonates in parallel with that capacitor at the frequency of interest. The impedance transformation capacitor and inductor, labeled C_{Trans} and L_{Trans} in Figure 3-5, perform the transformation. If the impedance of the load and the desired load impedance are known, the values of L_{Trans} and C_{Trans} can be determined analytically by knowing that the impedance seen by the device needs to be $R_{OPT} + j^*0$, i.e., the imaginary part of the impedance seen should be zero. This gives a two-equation, two-unknown system, providing an analytical solution for L_{Trans} and C_{Trans}.

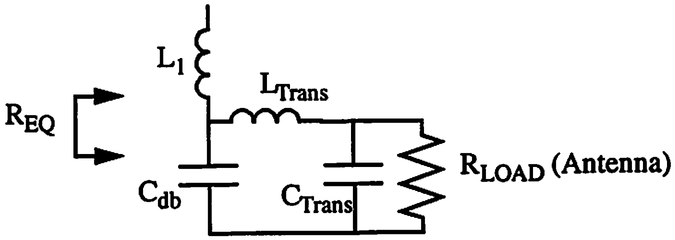


Figure 3-5. Output Network Structure

At this stage, assuming that the output capacitance of the device can be estimated, the inductance to be used in the LC tank at the output can then be easily determined. An initial estimate of the device size can be approximated using the current information obtained in Section 3.3.1, giving a first-order estimate of the passive components to be used in the output network.

3.3.3.2 The Drain Current

It was previously mentioned that the input gate overdrive (the V_{GS}-V_T) could be decomposed into its fourier components using Eq. 3-4 through Eq. 3-6. A problem arises, however, in how to apply the square-law transfer function to the decomposed signal when generating the drain current I_D; one of the important issues that has been repeated in this work is that the device does not act only as a square-law device, even if the focus is restricted to the region where the device is on. It will enter the linear region at some point as well, and accounting for that is one of the goals of this work. The square law relation can not be applied blindly. The difficulty arises in that the drain current is dependent on the drain voltage waveform while the device is in triode; however, it is precisely that drain current which determines the drain voltage waveform. Therefore, the key goal of the design methodology should be to construct the drain current waveform, from which the output voltage waveform as well as the key PA metrics (such as efficiency and harmonic distortion levels) can be obtained.

3.4 Class C Design Methodology

In Chapter 2, the current method for designing non-linear PAs was described, and the fact that the empirical nature of such a method was not conducive to PA design in integrated CMOS was discussed as well. As such, a less empirical method would be preferable, as the well-characterized models available when using discrete devices are generally not available for integrated CMOS devices. While the finalization of the design will require compute-intensive simulations, without a less-intensive method, finding a reasonable starting point with which to begin the design is not so easy.

In the previous section, the circuit topology was discussed and a method for determining a starting range for the particular components was described. Now that a starting point for the circuit itself is known, the determination of the drain current can proceed. However, this is a complex endeavor; as stated earlier, the drain current is dependent on both the input voltage (V_{GS}) and the output voltage (V_{DS}). But it is precisely the drain current that determines the output voltage! This interdependence makes a direct, closed-form solution an unreal goal.

The question then arises as to how to determine the output current. An analogy can be made between the case being analyzed in this work and the case where a particular value needs to be found for a particular variable in a similar system. For example, assume that the equation

$$x = Ke^{(x-1)} \qquad \text{Eq. 3-11}$$

needs to be solved for the variable x, where K is some known constant. A cōmmon method of solving this equation for the correct value of x is iteration, in which a starting estimate of x is used in the right side of the equation, and the resulting value is then assigned to x and then used in the right side of the equation again. Eventually, the result of the right side will level off and approach the desired value. The number of iterations required to approach the final value will depend on the initial estimate.

A similar method can be used in the case of determining the output current waveform. Certainly, the desired final goal is not a single value but a waveform; however, the ideas are comparable. In order to approach the final value of the drain current, an initial estimate must first be assigned. That initial estimate yields a particular drain current, which can then be used to refine the initial estimate of the drain current. Note that a key issue in determining the drain current is that the region can be in either the saturation or triode region while on. As the drain current waveform is refined with each iterative pass, so too is the drain voltage waveform, further correcting the transition region between the saturation and triode regions.

Furthermore, this method should be able to provide a solution for both the "true" Class-C PA and the mixed-mode Class-C PA, which were described in Chapter 2. The difference between the two cases is that in the case of the "true" Class-C PA, the device remains in the saturation region when on, whereas the mixed-mode PA enters the triode region. This is an extremely simple check to make, and allows for both "true" Class C and mixed-mode Class C PAs to be designed. As a result, the method for determining the drain current of the device used in the PA follows the steps illustrated in Figure 3-6.

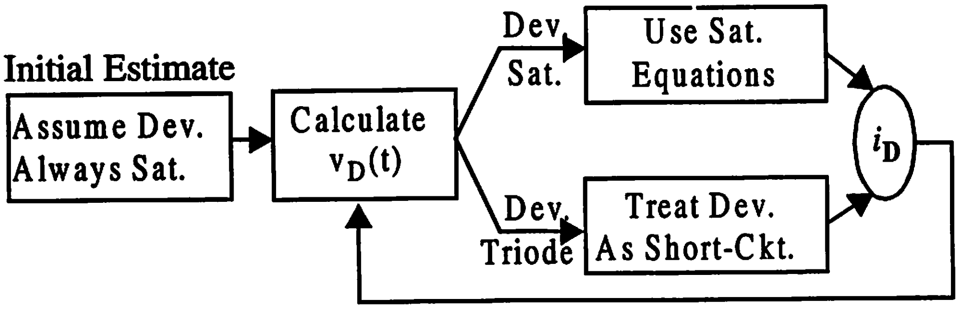


Figure 3-6. Steps for designing Integrated Class C PA

From the steps shown in Figure 3-6, the first task in this approach is to determine a viable and reasonable initial "value" of the drain current waveform, one that is simple to start with yet will reach the final value without too many iterations. The initial estimate should be a simply attained one, because otherwise, the iteration technique would be less useful. It should also allow an approximate final value to be reached without too many iterations; again, if the number of iterations needed is too large, the technique is less useful. The best method of determining the initial estimate is to examine what is known before starting the iteration, and using that data to make the estimate. In this case, the input signal is fixed (for a given bias point and signal amplitude), and the equations determining the operation of the device are known. One possible method of determining the initial value to be used in the iteration is to start by assuming that the device is in the saturation region while on, i.e. that it doesn't pass into triode. In this manner, the output waveform can be determined (even though it is known to be incorrect).

Once the initial estimate is available, the iterative procedure can begin. In each step, the drain current waveform needs to be determined, and a resulting output voltage waveform must also be determined. Once the output voltage waveform is known, that knowledge can be used in the next iteration to determine where the device transitions from the saturation region to the triode region, and a more accurate drain current waveform can be obtained. That process is then repeated until an approximate solution is reached.

3.4.1 The Initial Estimate

One possibility for the initial estimate for the drain current waveform, as stated above, is to assume the device is in saturation for the entire region, and then to determine the drain current waveform that results from that assumption. The drain current waveform, then, is given by Eq. 3-1 to be

$$i_D = \frac{\mu_n C_{OX}}{2}\frac{W}{L}(v_{GS} - V_T)^2 \qquad \textit{for } v_{GS} > V_T \text{ ,} \qquad \text{Eq. 3-12}$$

ignoring the $1{+}\lambda V_{DS}$ term as discussed previously, and 0 for all other times in the period of the input waveform. The radian values that mark the boundaries for which the input waveform is greater than the threshold voltage can be obtained from Eq. 3-9.

It is apparent that the output current predicted by this equation will be much larger than the actual output current, as the current through a device in the saturation region is significantly larger than the current through the device in the triode region. While the output current will be significantly large, the equation above provides a simple closed-form starting point to the analysis. Using Eq. 3-12, the shape of the output current is a pulsed current as shown in Figure 3-4 (the corresponding input voltage is shown in Figure 3-4 as well). The current can be broken down into its

Fourier components as described previously. Doing so gives the DC, first, second, and third harmonic components to be

$$b_{i_0} = \frac{k'}{2}\frac{W}{L}\frac{1}{2\pi}\left((V_B - V_T)^2 y + 4V_A(V_B - V_T)sin\frac{y}{2} + V_A^2\left(\frac{y}{2} + \frac{1}{2}sin y\right)\right) \qquad \text{Eq. 3-13}$$

$$a_{i_1} = \frac{k'}{2}\frac{W}{L}\frac{1}{\pi}\left(2(V_B - V_T)^2 sin\frac{y}{2} + V_A(V_B - V_T)(y + sin y) + \frac{2}{3}V_A^2 sin\frac{y}{2}\left(2 + \left(cos\frac{y}{2}\right)^2\right)\right) \qquad \text{Eq. 3-14}$$

$$b_{i_2} = \frac{k'}{2}\frac{W}{L}\frac{1}{\pi}\left(-(V_B - V_T)^2 sin y - 2V_A(V_B - V_T)\left(sin\frac{y}{2} + \frac{1}{3}sin\frac{3y}{2}\right) - \frac{V_A^2}{2}\left(sin y + \frac{y}{2} + \frac{sin 2y}{4}\right)\right) \qquad \text{Eq. 3-15}$$

$$a_{i_3} = \frac{k'}{2}\frac{W}{L}\frac{1}{\pi}\left(\left(-\frac{2(V_B - V_T)^2}{3}sin\frac{3y}{2}\right) + \left(2V_A(V_B - V_T)\left(-\frac{sin y}{2} - \frac{1}{4}sin 2y\right)\right) - V_A^2\left(\frac{2}{3}sin\frac{3y}{2} + sin\frac{y}{2} + \frac{1}{5}sin\frac{5y}{2}\right)\right) \qquad \text{Eq. 3-16}$$

The variables V_B, V_A, and y are all defined as represented in Figure 3-4.

3.4.2 Iterating to reach final value

The equations above can then be used to estimate the resulting output voltage waveform. Since the output network is known, the approximate output voltage is just determined by the current flowing into that network. Approximating the output current using the first three harmonic components, the output voltage is

$$V_O = V_{DD} - [a_{i_1} sin(\omega t) z_O(\omega) + b_{i_2} cos(2\omega t) z_O(2\omega) + a_{i_3} sin(3\omega t) z_O(3\omega)] \qquad \text{Eq. 3-17}$$

The output voltage given by using the first estimate of the current will more than likely be quite unrealistic, hence the need for the iteration. With the saturation-region assumption, the current will be severely over-estimated, leading to the resulting output voltage swing being significantly larger than it is in reality.

However, the over-estimated current has provided an initial output voltage waveform, which can then be used to refine the prediction of the output current. Once the output voltage is known, the approximate regions where the device transitions from saturation to triode can be predicted, and the full first-order MOS model can be used to obtain the next iteration of the output current.

While in the triode region, the MOS device acts like a variable resistor, and as such, the current through the device is really determined by the other elements of the circuit. Those other elements of the circuit are primarily the inductors and capacitors that make up the output network. If it is assumed that the device in the triode region has an extremely small on-resistance, the voltage across the inductor to the supply will have an approximately constant voltage across it. A simple approximation in this case is that the device is acting approximately like a virtual short-circuit; it is holding the voltage across itself virtually constant (due to the extremely small on-resistance). While this is a decided approximation, it simplifies the analysis and allows the methodology to move forward. Making this assumption, the current through the inductor connected to the supply will ramp up linearly during the period when the device is in triode (as it has an approximately constant voltage across it). The voltage across the parasitic drain-bulk capacitor at that time would similarly be approximately constant, and therefore no current flows through that capacitor. Finally, if the assumption is made that the current through the load is ideally the fundamental component of the drain current, it is straightforward to determine the current through the device. The sum of the currents leaving the output node must sum to zero; in other words,

$$i_D + i_{L_{VDD}} + i_{C_{DB}} + i_{OUT} = 0 \Rightarrow i_D \approx -(i_{L_1} + i_{OUT}) \quad . \qquad \text{Eq. 3-18}$$

With the above information, the drain current during the period when the device is in the triode region can be ascertained. In a given iteration, i_{OUT} would simply be the fundamental component of the current of the drain current from the previous iteration, and i_{L1} is linearly ramping during the period that the device is in triode.

Putting the piecewise approximation together gives a description of the drain current i_{D} which can be broken down into its fourier components again to start the second iteration. Due to the complexity of the drain current representation, the use of a mathematical tool able to perform simple integration is beneficial. Obtaining a closed-form equation for the drain current that can then be used to obtain Fourier coefficients would be prohibitively complex. A simple tool such as MathCad(TM) can perform these definite-integrals in a straightforward manner, and the results can be used in the next stage of the iteration.

Once the iteration converges, it is quite straightforward to calculate some of the key metrics, including the efficiency of the PA. The efficiency is simply the amount of power delivered to the load divided by the DC power consumed. The DC power is simply the supply voltage multiplied by the DC component of the current. The power delivered to the load is determined by the amount of current that flows through the load. This simple procedure can be repeated with the free design variables swept across certain ranges in order to generate an first-pass solution, which can then be used in a circuit simulator in order to optimize the design. In order to demonstrate this design methodology, a design method is given in the next section.

3.4.3 A Design Example

The basic design methodology has been described above. The easiest way to see the use and benefit of this method is to go through a design example. In a simple design example, assume that a 125-mW, 1.89-GHz PA must be designed. The design methodology presented above can now be used to determine an approximate circuit implementation. In this design example, any CMOS limitations presented in the following chapter will be ignored for the time being. While these limitations are critical, methods of limiting their impact will be presented in Chapter 5, and as a result, the approximate solution can be determined independent of those limitations.

At this stage, the general circuit topology and output requirements are known. The unknown which most impacts the final circuit is the size of the transistor to be used in the PA. The device size impacts the output capacitance as well as the amount of current available for a given input voltage swing. It may also indirectly affect the optimum output resistance, for the peak current produced varies with the device size.

In a design of this type, a range of device sizes will be able to generate the output power required by the system. Each device size will have a corresponding bias and input signal amplitude, as well as output network. In the end, the one that appears to be most optimum will generally depend on outside factors. For example, if reducing the magnitude of the input signal to this final stage is required, a larger device which requires less input signal to provide the same current may be the more optimum solution. However, if the output device must be limited because of other factors (generating the output matching network, for example), a different output implementation may be the optimum one.

As stated earlier, the degrees of freedom available to the designer include the size of the device (a theoretically infinite range of possibilities), the biasing of the input and the magnitude of the signal applied there (which are generally related). In applying the iterative design procedure to the problem of designing a class-C PA, one of the first key steps is to identify a range of potential device sizes and input signal magnitudes which may satisfy the design requirements. A simple lower bound on the potential range of values can be set by examining the magnitude of the fundamental frequency component of the drain current if the device is always saturated (the assumption that provides the initial estimate in the design method). The amount of current provided by the mixed-mode PA will be significantly less than that provided by the true-mode PA, since the device will enter the triode region where the amount of amplification is reduced drastically. Furthermore, if it is assumed that the device will be in the triode region for an extended period, the fundamental frequency component of the current while in saturation should be somewhat larger than the amount required to provide the peak power mandated by the specifications.

Conversely, a reasonable upper bound will likely be set by concerns related to the efficiency and the ability to adequately drive the device. The current consumed in the prior stages required to drive the final output stage is completely wasted, in the sense that none of that power is consumed in the load. Only the final output stage will transfer power to the load. The upper bound is dependent on the particulars of the design, including the technology used, and the desired efficiency. For technologies with the ability to provide large amounts of gain, the upper bound on the device size might be quite large, as the ability to drive a large load may be available, and vice versa.

In this design example, the potential range of device sizes used was 1000μm to 4000μm, while the peak signal input magnitude used was 2.1V. Using the initial estimate that the device is in saturation, the fundamental frequency component of the output current varies as shown in Figure 3-7(a) and Figure 3-7(b). The output power increases quadratically, as would be expected from the use of the simple square-law model described in Section 3.2. However, the efficiency rises without bound as well, which is clearly not realistic. Furthermore, the magnitude of the output voltage (for a given, fixed output resistance) also increases without bound, which will clearly send the device into the triode region for some portion of the cycle.

As a result, the iterative procedure is required in order to accurately predict the operation of the PA. In this procedure, it is quite simple to vary the design variables and note the affect on the output power. After a few iterations, the output power and

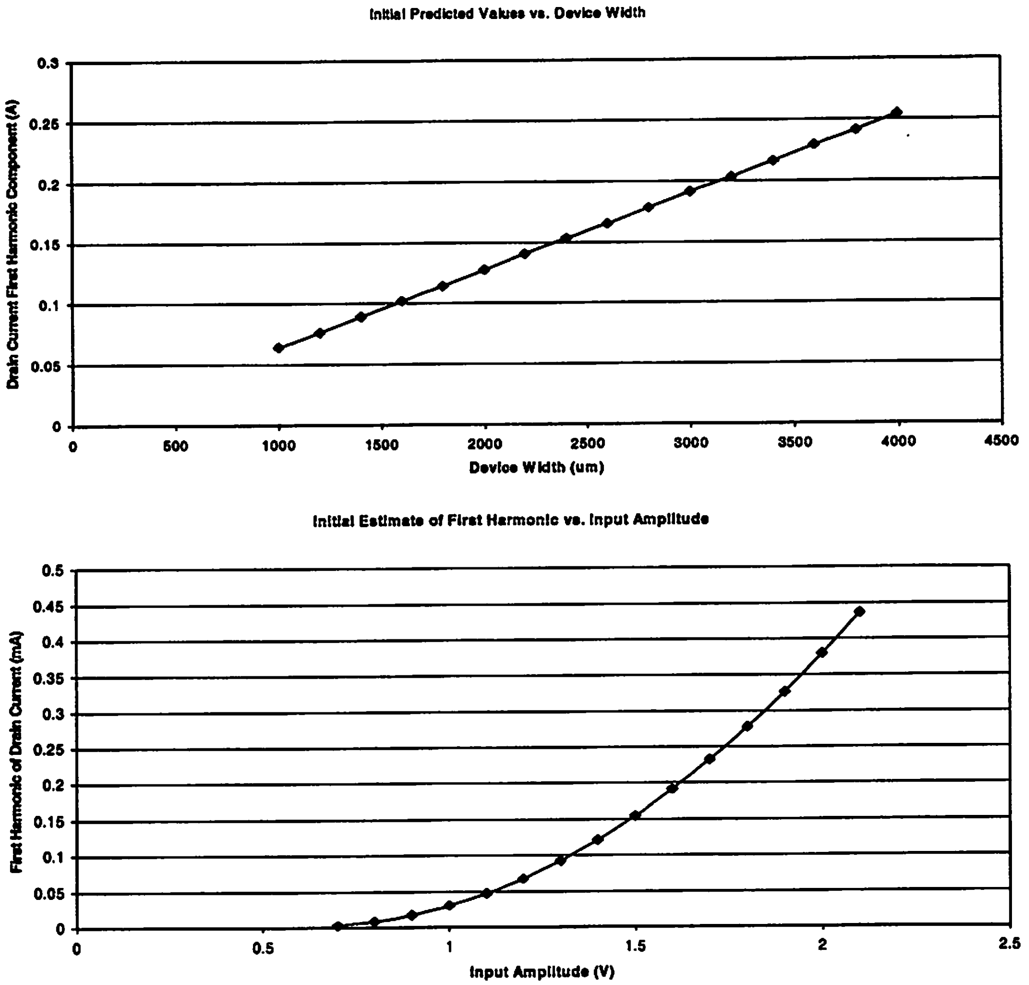
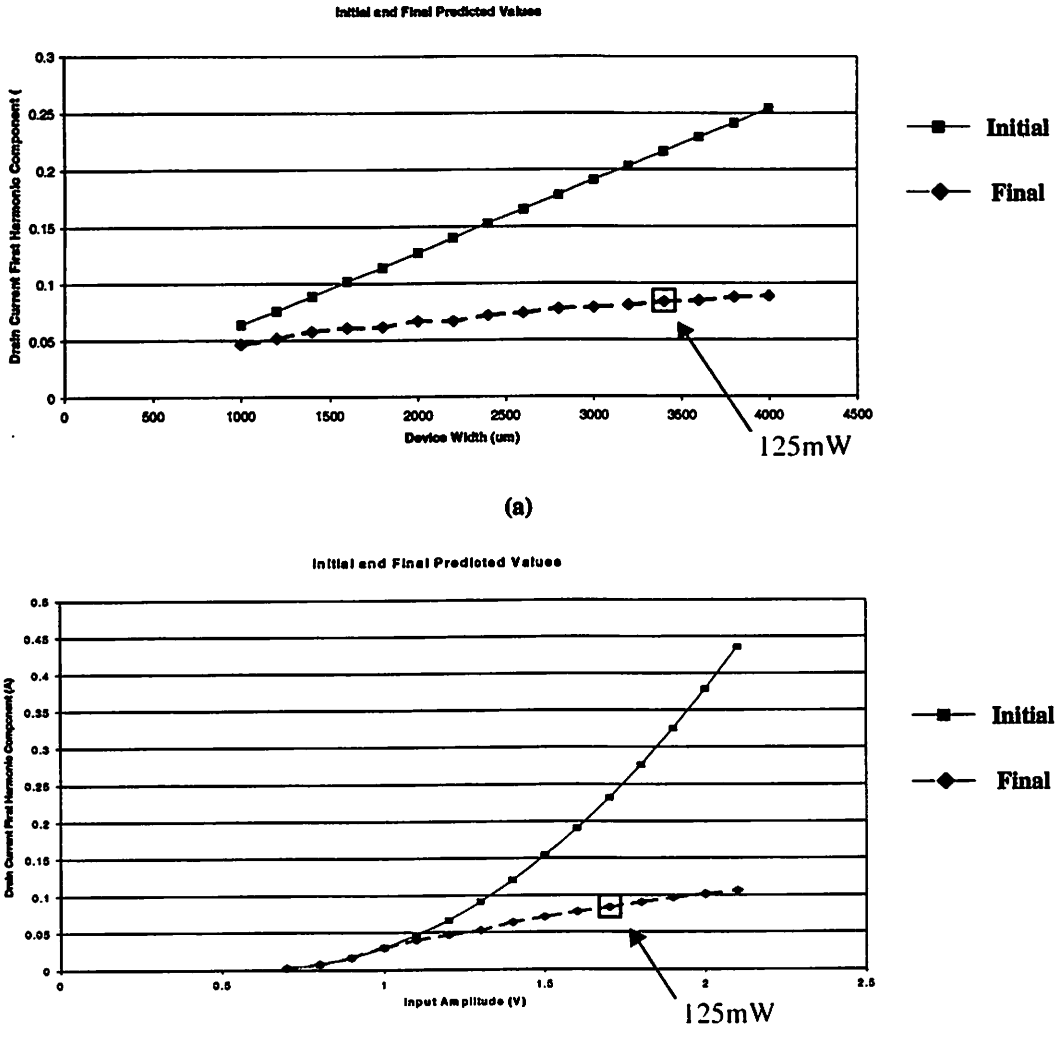


Figure 3-7. Initial estimate plots versus varying width and input amplitude

efficiency can be plotted against the width and the input signal amplitude, as shown in Figure 3-8. For example, one solution with a reasonable input magnitude is highlighted on the figures. For the input amplitude being approximately 1.7V, a device size of approximately 3,000μm/0.35μm will provide the required output power; similarly, a device of size 3,400μm/0.35μm with an input amplitude of 1.6V will also provide the required output power. If more input amplitude can be provided, a smaller device may potentially be used. Conversely, if that amplitude cannot be provided by the previous stages, a larger device might be required in order to provide the same level of current.

Furthermore, the effect of changing the input bias can be seen on both the output power and the harmonic components of the output. In Figure 3-9, the plots shown show that the level of the 3rd harmonic in the output current varies as the bias



(a)

(b)

Figure 3-8. Drain Current Fundamental Component Initial and final estimates versus (a) device width, and (b) input signal amplitude.

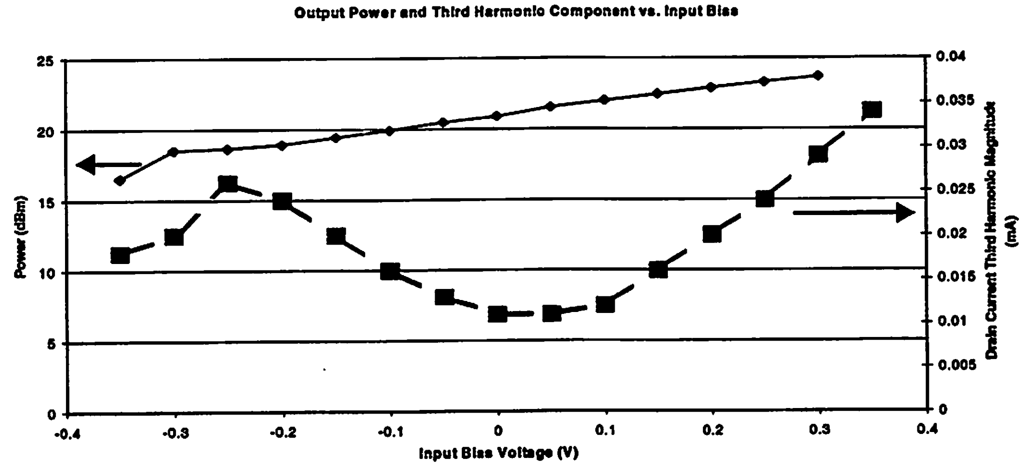


Figure 3-9. POUT and Harmonic Levels versus Input Bias

point is reduced, which is straightforward to understand. As the window of time in which the current flows is shortened, the peak current must be larger in order to produce the required magnitude of current at the fundamental frequency. However, the amount of time for which the device is in saturation will vary as well. As a result, the higher-order harmonic components will be changing, and that change is reflected in the graph in Figure 3-9.

Finally, in order to investigate the validity of this method, the potential circuit was simulated in HSPICE, and in Figure 3-10 the results are plotted against the results obtained from the design method described in this chapter. As can be seen, the results match very well, and allow the designer to get an in-depth understanding of the operation of the circuit.

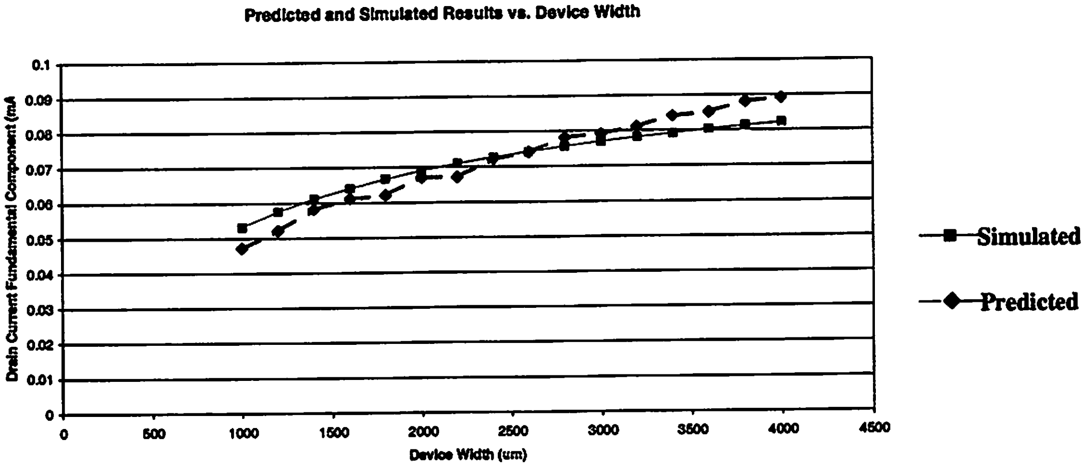


Figure 3-10. HSPICE results versus MATHCAD results

3.5 Earlier Stages

A similar design procedure can be used for the earlier stages in the signal path. Each stage can be modeled independently using one device and an output matching network, although in the case of the previous stages, those matching networks will generally by purely reactive instead of resistive. Furthermore, because the previous stages will not require the level of overdrive that the final, power stage requires, the previous stages will generally consist of devices that remain in saturation throughout the portion of the cycle that they are conducting (depending on the class of the previous stages). By designing the final output stage first, the amount of swing required at the output of each prior stage will be known prior to designing the stage, providing a design goal for each stage.

3.6 Conclusion

In this chapter, a simple design methodology which approximates a simple hand-analysis for the design of Class-C PAs was presented, which is useful in designing PAs when a fully-characterized empirical model (such as those available for the discrete transistors commonly used in PAs) is not available. This design method uses the most simple MOS circuit model, the square-law model. This method was used in a simple design example to provide a glimpse into the procedure by which the design method can be used to design the PA. The result of the design method can then be used in SPICE with a more complicated model and with more of the "real-world" encumbrances to refine and finalize the design.

4

CMOS Technology Limits

4.1 Introduction

In order to implement the PA in a CMOS technology, one must first understand the limitations of CMOS with respect to the requirements of the PA. The PA must deliver a large amount of power to the antenna while consuming a minimum amount of power. In general, in order to maximize the Power Added Efficiency (PAE), which was explained in Section 2.3, the magnitude of the input signal should also be quite small. The smaller the size of the input, the less power consumed by the previous stage in providing the required signal drive. So the gist of this is that the PA must provide considerable gain without dissipating a great deal of power.

There is another consideration that must be accounted for in the design process, dealing with what is often thought of as a second-order effect: namely, parasitic resistance. Because of the large amounts of power that must be delivered to the load, it follows that a large amount of current is required. While parasitic resistances throughout the circuit are generally an afterthought in many designs, the power dissipated in the parasitic resistance can be significant when calculating the efficiency. Wherever possible, the current must be minimized, but especially in the output stage of the PA, where the most current will be flowing.

However, CMOS technologies are, in general, not conducive to satisfying the above requirements. The next few sections will describe some of the limitations of CMOS technologies, especially some of the advanced, sub-micron technologies, with regards to the implementation of a relatively high-output power PA for mobile cellular applications.

4.2 Breakdown Voltage In Sub-Micron CMOS

The advent of sub-micron CMOS processes has really fueled the investigation into implementing high-performance RF circuits in CMOS processes. The f_T of a MOS device, which is the maximum frequency at which current gain is possible, is given by

$$f_T = \frac{g_m}{C_{GS}} = \frac{\mu_n C_{OX}\frac{W}{L}(V_{GS} - V_T)}{WLC_{OX}} \propto \frac{1}{L^2} [12]. \qquad \text{Eq. 4-1}$$

So as the technology, measured by the minimum MOS device length (L), gets better, the f_T of the device improves. As the f_T improves, the gain at the RF frequencies of interest in cellular applications increases, and CMOS starts to become a viable technology for some of these functions. However, in the case of the PA, while the increased gain is a benefit, there is a significant drawback that is associated with the decreasing feature size: the issue of oxide breakdown.

The gate oxide material used in the MOS device supports an electric field, which is generated by the voltage across the oxide. In general, oxides that are not significantly defective are thought to break down in two different ways [22]. In the first mode, the oxide is thought to be defect-free, and can support electric fields on the order of 8-12 MV/cm. The second mode of breakdown occurs at fields of moderate strength, on the order of 2-6 MV/cm. In this mode, the application of these moderate fields can reduce the lifetime of the device, as the long-term stress across the oxide causes it to breakdown either instantaneously or earlier than expected (known as time-dependent dielectric breakdown - TDDB). It is thought that the existence of defects in the oxide leads to the reduction in the lifetime of the device.

The exact causes for the breakdown and their mechanisms are not known for certain; however, there are several theories which are highly plausible. One of the most widely accepted theories is the hole generation and trapping model[23][24], in which electrons under a sufficiently large field are injected into the conduction band of the oxide by Fowler-Nordheim tunneling, as shown in Figure 4-1[22]. Once in the oxide conduction band, these electrons are accelerated by the large field toward the MOS substrate (which serves as the capacitor anode in the case of an NMOS device). A small percentage of these electrons generates electron-hole pairs in the oxide. A fraction of the generated holes may then be trapped in oxide traps (localized areas which are more likely to trap positive charges than normal). These locally trapped positive charges increase the field locally, increasing the tunneling current density. This positive feedback effect causes the localized positive charge to increase. Once the density of that localized charge reaches a critical level, the tunneling current through that region is increased, and breakdown occurs.

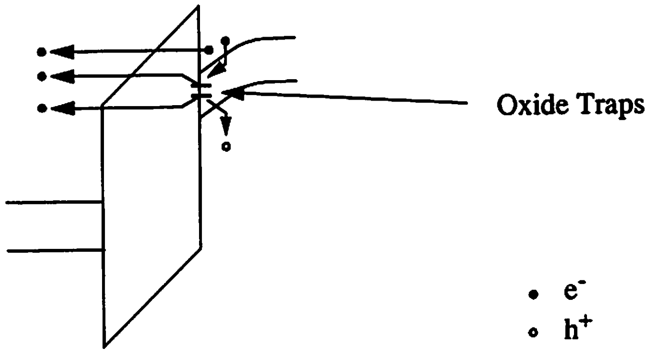


Figure 4-1. Fowler Nordheim Tunneling

Another model of the breakdown mechanism agrees that electrons are injected into the oxide through Fowler-Nordheim tunneling, but then postulates that these electrons then undergo acceleration in the oxide[25]. The energy the electrons acquire in the acceleration is dissipated when the electrons reach the anode by creating atomic defects (for example, breaking bonds between silicon and oxygen) at the oxide interface. These defects are positively charged, and thus attract any newly injected electrons. As more electrons are injected, the defects directs the growth of the defects into the dielectric; over time, a conductive path of these defects can be created between the gate and the drain, eventually causing the breakdown of the oxide.

While the exact cause for oxide breakdown is not known, and the existing theories do not necessarily agree, empirically, the progress of oxide breakdown can be predicted. A general equation which can predict the lifetime of a device is

$$t_{BD} = \tau_0(T)e^{\left[\frac{G(T)}{E_{ox}}\right]}, \qquad \text{Eq. 4-2}$$

where τ_0(T) and G(T) are temperature dependent factors. At 300K (room temperature), the equation simplifies to

$$t_{BD} = 10^{-11}e^{\left[\frac{350t_{ox}}{V_{ox}}\right]}, \qquad \text{Eq. 4-3}$$

where t_{OX} is in centimeters, V_{OX} is in volts, and t_{BD} is given in seconds.

Furthermore, the strain on the oxide differs if the voltage is constant (i.e. a DC signal) or time-varying (an AC signal). An analysis of the impact of time-varying signals gives[26]

$$1 = \frac{1}{\tau}\int_0^{t_{BD}} exp\left(\frac{-GX_{eff}}{V_{OX}(t)}\right)dt, \qquad \text{Eq. 4-4}$$

where τ and G are constants and X_{eff} is the effective oxide thickness. The effective oxide thickness accounts for the severity of defects in the oxide; for example, in a defect-free oxide, X_{eff} would just be the physical oxide thickness t_{OX}. Eq. 4-4 shows that it is the accumulation of strain over time that causes the degradation in the long-term reliability of the device under strain. In the case of sinusoidal signals, the stress on the device would be considerably less than it would be in the case of constant DC voltage. However, it is generally much easier to work with Eq. 4-3 than Eq. 4-4, and thus it is preferable to assume a constant stress on the device. Furthermore, because long-term reliability is such a critical issue, it is often more wise to err on the side of safety than to try to push the boundaries of the device; a device which breaks down quite easily under stresses that are designed to be standard is of no use to anyone.

In general, the basic idea presented in Eq. 4-3 is that the oxide can support electric fields up until a certain magnitude, labelled in this work as E_{bd}. Once the electric field in the oxide is greater than that critical field, the oxide will eventually break down and will no longer function as a insulating gate; in fact, gate current will flow across the nodes which caused the field to exceed the E_{bd}. The electric field in the oxide is related to the voltage across the oxide by the thickness of the oxide, i.e.

$$E_{OX} = \frac{V_{OX}}{t_{OX}} \qquad \text{Eq. 4-5}$$

So, in order to keep the electric field in the oxide below a maximum, critical value, the voltage difference across the oxide must be kept below the corresponding maximum voltage. The cross-section of an MOS device is shown in Figure 4-2. The voltages V_{GS} and V_{GD} must be less than the voltage Vcrit (or tox*Eox), or the oxide will break down, rendering the transistor useless.

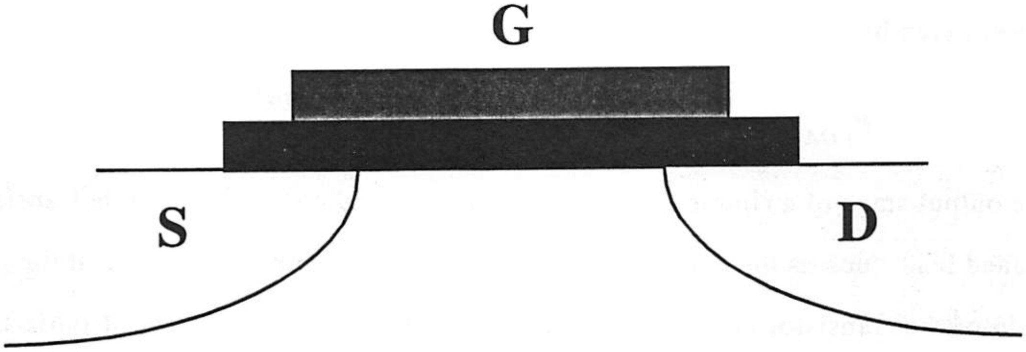


Figure 4-2. MOS Device Cross-Section

As the minimum length in a CMOS process decreases, so too does the thickness of the gate oxide; in order to ensure that the gate voltage retains control of the channel, the oxide thickness must be reduced. Otherwise, the transistor is subject to severe short-channel effects, such as Drain-Induced Barrier Lowering (DIBL)[22]. In essence, if the oxide thickness is not decreased, the portions of the channel proximate to the drain will be controlled more by the drain than the gate, resulting in the barrier to conduction, i.e. the threshold voltage V_T, becoming dependent on the drain voltage. In order to prevent this from occurring, the oxide thickness is reduced as the channel length is reduced.

The question arises, why is this breakdown voltage issue so important in the design of the PA? The answer lies in the issue of how to generate a certain amount of power when the available voltage swing is limited. The trade-off is to generate more current, and this current is a significant issue in the overall efficiency of the PA. The current is an issue not only because of the amount of current drawn from the supply, but because of current flowing through parasitic resistances and increasing the overall power dissipation of the circuit itself. If it is assumed that the output load can be modeled by a given resistance (often 50Ω in the case of RF system implementations), then a single-frequency, constant magnitude signal across the resistance will deliver a power given by

$$P_{LOAD} = \frac{\bar{V}_{LOAD}\bar{I}_{LOAD}}{2} = \frac{\bar{V}_{LOAD}^{2}}{2R_{LOAD}} = \frac{\bar{I}_{LOAD}^{2}R_{LOAD}}{2} \qquad \text{Eq. 4-6}$$

The output stage of a simple PA implementation would consist of a single transistor and a tuned load, such as the one shown in Figure 2-7. The terminal voltages at the gate and drain of the transistor are presented again, along with the difference of those terminal voltages, in Figure 4-3. This difference voltage is the oxide voltage, which must be kept below the critical oxide voltage in order for the PA to remain functioning properly. The fact that the MOS device is an inverting device means that the maximum strain on the oxide occurs when the input voltage is at its minimum value (and thus the output

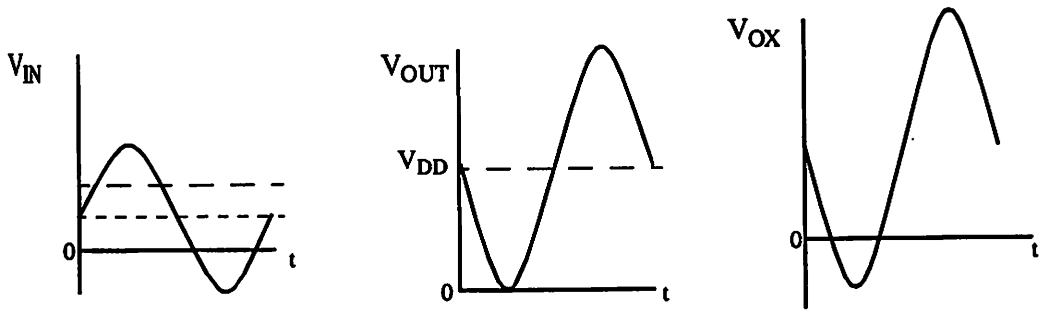


Figure 4-3. PA Waveforms and Oxide Voltage

voltage is at its maximum value). In the case of a Class C implementation, because the input is biased below the threshold voltage of the device, it is possible for the input voltage to go *negative*. So the maximum absolute voltage on the drain can be severely limited. That being the case, there is only one way to still deliver a given amount of power into the load: the reduced voltage must be traded for an increased current, and the effective load impedance that the device sees must be reduced as well, so that the maximum voltage swing across the load does not overstress the gate oxide. Eq. 4-6 shows that in order to achieve a given amount of power with a limit on the available voltage, it is necessary to increase the current to the appropriate level.

Earlier, however, it was made clear that drawing a great deal of current will detract considerably from the efficiency. It is in this area that processes like Silicon bipolar and Gallium Arsenide (GaAs) have an enormous edge over Silicon CMOS. Other technologies often have better frequency performance than CMOS; basically, they have a higher f_T for a similar feature-size device. Therefore, such state-of-the-art processes as are needed in the case of CMOS are not required for operation at a given frequency. As a result, the breakdown voltages are considerably higher, requiring less current drawn from the supply and increasing the overall efficiency. As will be seen in later sections, those processes also have other advantages which benefit PA implementations. Not only is this breakdown issue directly a cause for concern, since the current requirement can go up, quite dramatically in cases, but it also impacts other issues indirectly, as will be discussed in greater detail.

4.3 CMOS Transconductance

In the previous section, it was established that considerable current is required in a CMOS PA in order to deliver a given power to the load. Knowing that, how is the current generated? That is, what is the mechanism by which that current will be delivered to the load? Simply stated, the MOS transistors available in the CMOS process are used as transconductance devices; the transistor, in response to an input voltage signal, supplies a current signal. A simple first-order approximation of the MOS device, as used in Chapter 3, says that the large signal output current is related to the input voltage by

$$I_{OUT} = \frac{1}{2}\mu_n C_{OX}\frac{W}{L}(V_{GS} - V_T)^2(1 + \lambda V_{DS}). \qquad \text{Eq. 4-7}$$

The MOS device is a square-law device with respect to its input signal drive, V_{GS}. Moreover, that equation is only good in the region where both the V_{GS} is larger then the threshold voltage *and* the V_{GS} is larger than the V_{DS}. In this area, the silicon MOS device falls short when compared to silicon bipolar or GaAs devices. In the bipolar device, the relationship between the output current and the input voltage is an exponential one, not a square-law. The result is that for a given overdrive voltage, bipolar devices can provide significantly more current than a CMOS device. Moreover, bipolar devices remain in their high-amplification region for a longer period than MOS devices, since the transition out of that region occurs (to first-order) at a fixed voltage (namely, when the base-collector region starts to be heavily forward-biased [12]). In GaAs, the devices used are often FETs, similar to what is used in silicon; however, GaAs has a significantly higher mobility (μ_n in Eq. 4-7) than silicon, so the amount of current generated for a given input overdrive is also higher.

The fact that silicon MOS is a poor current-drive technology leads to two potential solutions, both of which have their pitfalls: first, the signal drive amplitude can be increased in order to accommodate the smaller transconductance, or, second, the device size can be increased in order to generate more current. Increasing the amplitude of the input signal requires the previous stage, i.e. the one providing the input signal, to consume more power in creating the input signal. As stated in Chapter 2, power consumed by previous stages in driving the final output stage is entirely "wasted": that is, it directly adds to the power consumption, but does not directly add to the power delivered to the load. Unfortunately, keeping the input amplitude constant while increasing the device size has virtually the same effect. Increasing the device size causes the capacitive load to the previous stage to increase. While driving an ideal capacitive load requires no real power to be consumed, in reality, increasing the capacitive load causes more power consumption. The parasitic resistances in every circuit path alone lead to more power dissipation if the current drive is increased. Both of these potential solutions reduces the efficiency of the CMOS PA implementation. In reality, more than likely a combination of the two solutions, increasing both the device size and input signal amplitude, would be used.

4.4 Current Consumption

There is another significant reason why current consumption is so critical in the design of a PA for a mobile wireless device, and that is the fact that the battery in a mobile wireless device is a reservoir of charge. The battery has a limited amount of charge, and that charge, when it flows into the chip, is the current used by the device. Therefore, the more current that the circuit requires, the shorter the battery life will be. The PA, which for a cellular communications system will more than likely consume the largest amount of current of any of the analog functional blocks, will drain that battery significantly. If the voltage swing is limited, and the current consumption must be increased in order to deliver a given amount of power, the battery life is reduced. If, furthermore, more current must be consumed in the pre-amplification stages in order to deliver that same amount of power, the battery life is reduced still.

The issue of battery life is a significant one, as it impacts the long-term cost of the wireless device. Since one of the goals of this work is to explore the impact of integration as it pertains to both the cost and size of the device, this battery life reduction is a critical component of the overall performance and attractiveness of the end result. A small reduction of the efficiency or other performance metrics of the PA when implemented in CMOS may be allowable if there were little reduction in the overall performance of the system; that is, if the PA is slightly less efficient, but the system as a whole consumed a comparable amount of power (or current), it would still be an attractive solution. However, if the PA consumes so much current that the overall system performance suffers, then the integrated CMOS implementation is not a very viable one.

It should be noted that while this is *not* a limitation directly inherent to the implementation of the PA in CMOS, it does result directly from the two issues described in the previous two sections, which are limitations inherent to CMOS processes. As such, this limitation does derive from implementing the PA in CMOS, and must be addressed.

4.5 Output Stage Device Sizes

The size of the output stage devices is a critical issue in the design of any PA, as that size will be a determining factor in how much power is consumed in the previous stages. Generally, in order to deliver a large amount of power to a load, a PA will be designed as a multi-stage amplifier. The stage preceding the PA, in the case of a RF transmitter, is usually a mixer or modulator, which will only be able to provide a small signal drive. The first stages of the PA will simply be used to supply enough signal drive to the output stage, also known as the power stage. As stated earlier, the power consumed in these first stages is entirely "wasted,", i.e. none of it is being delivered to the load.

The CMOS limitations raised in the previous two sections greatly impact the sizing of the output stage of the PA. Section 4.2 explained that because of oxide breakdown considerations, the amount of current that the PA needs to deliver must be increased to compensate for the limited voltage swing that is available. That alone causes the output stage devices to be sized up. Furthermore, Section 4.3 shows that the MOS devices in silicon CMOS have poor transconductance; in order to generate a given current, larger signal drives or larger devices, or both, are required. Both factors directly add to the need for larger devices.

The fact that larger devices are needed is not such a detriment if only the output stage is examined. However, when the need to adequately drive the input is considered, the size of the output devices becomes a significant issue indeed. Assuming that the gate impedance of the MOS transistor can be modeled as a simple capacitance C_{GS}, the amount of current needed to generate a voltage swing of a given magnitude can be determined by

$$\bar{I} = \frac{V}{Z_{in}} = \bar{V}j\omega C_{GS}. \qquad \text{Eq. 4-8}$$

As the magnitude of the gate capacitance increases, so does the current required. While ideal capacitances do not dissipate power, the power drawn from the supply in the case of the gate capacitance is actually dumped to ground, and not returned to the supply, so that current symbolizes energy drained from the battery source.

Unfortunately, this is not the only problem caused by the increasing device size. In these CMOS RF circuit implementations, the frequencies of interest are often quite close to the device f_T, i.e. the frequency at which the current gain of the device falls to one. The current gain at the frequencies of interest is significantly less than the

MOS current gain at low frequencies. In narrowband RF circuits, however, a method commonly used to counter the reduced gain at RF is to use inductors to tune out the capacitances in the circuit and create a peaked impedance. The use of a parallel resonant circuit dramatically reduces the gain at low frequencies, but since the operation of the circuit at those low frequencies is unimportant, inductive tuning increases the gain at frequencies approaching the f_T of the device. The basic effect of the parallel L-C Circuit Can be seen in Figure 4-4.

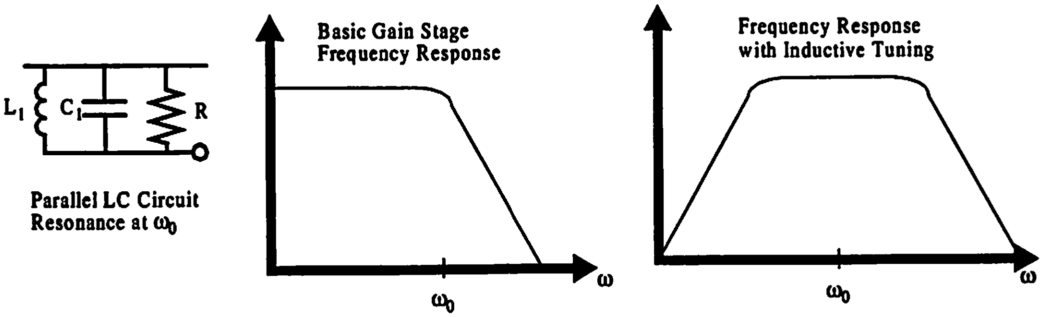


Figure 4-4. Parallel L-C Circuits

The inductor values that can be practically realized in single-chip integrated circuit systems are not without limits, however. In order to be useful, the inductor to be used must have a reasonable quality factor (Q), and it should be repeatable and somewhat immune to parasitics and process variation. The Q of an inductor is a measure of its ability to store energy; one simple metric for evaluating the Q is the ratio of the portion of the inductor's impedance which is imaginary to that portion which is real. An ideal inductor has an impedance which is entirely imaginary; it therefore has an infinite Q. However, real inductors are all subject to parasitic resistances, and so if a real inductor can be represented by an ideal inductor of value L and a resistance in series with that inductor of value R, then the Q of the real inductor can be calculated to be

$$Q = \frac{\textit{imaginary part of impedance}}{\textit{real part of impedance}} = \frac{\omega L}{R}. \qquad \text{Eq. 4-9}$$

In integrated circuit implementations, several factors work to limit the quality factor of the available inductors. Planar spiral inductors are the most common on-chip inductors used in integrated circuits. In general, these inductors are created by laying out a strip of metal in a square spiral structure, as shown in Figure 4-5. The inductor is laid out over the substrate of the integrated circuit technology. Each of the strips of wire has its own self inductance, and a magnetic field is created by the circulating current in order to increase that inductance[27]. However, the practical, usable values of inductance realized by using one of these structures is limited by several factors. First and foremost, the metal used in standard CMOS processes has a finite non-zero sheet resistance, and as the physical size of the inductor increases, this increase in the series resistance reduces the Q of the inductor. Newer processes with special metal layers (including both very thick Aluminum and low-resistivity Copper) are being investigated[28][29], but those specialized layers are not often available in standard CMOS processes. Secondly, as the inductor increases in physical size, the capacitance from the metal layer to the substrate increases. This capacitance sits in parallel with the inductor, creating a parallel tank circuit, without even accounting for the capacitor in the external circuit that the inductor is trying to tune out. The parallel tank circuit of the inductor structure has its own self-resonant frequency (SRF); above that frequency, the planar structure looks capacitive, and will be of no help in tuning out other capacitors. Ideally, the SRF should be significantly higher than the frequency at which the inductor will resonate with the external capacitance. Furthermore, the capacitance

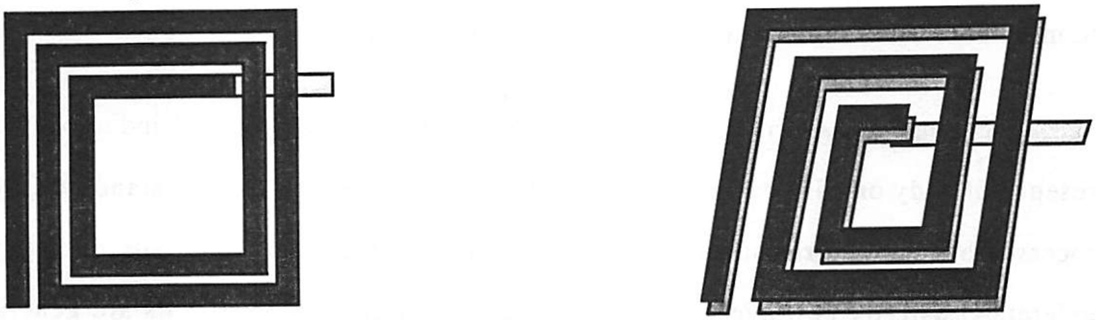


Figure 4-5. Spiral Structure of Integrated Circuit On-Chip Inductors

to the substrate also sees the resistance of the substrate; at frequencies well below the SRF of the planar inductor, that resistance can be neglected, but close to the SRF, those resistances will diminish the Q of the inductor still. All these factors combine to limit the maximum usable inductor values to on the order of 10nH.

Furthermore, a critical factor limiting the Q of on-chip spiral inductors is the presence of eddy or mirror currents in the substrate. In general, today's standard CMOS processes have a low-resistivity substrate (around 0.01 Ω-cm) which sits on top of a moderate-resistivity epi layer (around 10 Ω-cm). These mirror currents are generated by the current flowing through the inductor; the vertical magnetic field flowing through the inductor creates an opposite current flowing in the substrate. When that current exists in the low-resistivity substrate layer, the dissipated power reduces the Q of the spiral inductor structure. While some of the other non-idealities mentioned earlier (series resistance, self-resonant frequency) can be mitigated through thick metals or by increasing the distance between the top metal layer and the substrate, the only way to significantly reduce the impact of the mirror currents on the inductor Q is to use a wafer in which the low-resistivity substrate layer is not included. That is, the substrate below the inductor should either be extremely low resistivity (much lower than the 0.01Ω-cm common in today's CMOS processes), in which case there is no dissipated power due to any mirror currents, or very high resistivity, in which case the vertical magnetic fields would not be strong enough to generate any significant mirror currents. However, those wafers are not currently used in "standard" CMOS processes; they are primarily used in BiCMOS or bipolar processes. It is possible that special RF CMOS processes will use wafers without these epi layers; however, they are not used in digital CMOS processes currently.

On the other hand, the minimum inductor values are generally limited by parasitics. The previously mentioned concerns all place upper bounds on the on-chip inductor values; it would appear that the intrinsic Q of very small-value planar inductors would be reasonable. However, while that is true, the effect of parasitic resistances and inductances would then start to dominate. For example, assume a planar inductor structure was designed with an inductance of 0.05 nH and a Q of 5 at 2 GHz. The effective series resistance of that inductor would be 0.06Ω. Parasitic resistances throughout the circuit could well dominate that number so that the effective Q was significantly lower. Moreover, straight metal traces do have their own self-inductance; for inductors of that size, parasitic inductances might well alter the effective inductance that is seen in the circuit. As a result, reasonable on-chip inductors are limited to minimum of a few tenths of a nanoHenry.

This bound on the inductor values means that if the output stage devices are too large, it is possible that the inductors required to tune out the gate capacitances of the devices may not be realizable in CMOS. The previous sections detail limitations in CMOS that significantly drive the size of the output stages up. But if the required inductive tuning can not be realized, the ability to implement the PA is compromised.

4.6 CMOS Substrate Issues

One final aspect of CMOS technology must be considered, relating to full-scale integration. The "Holy Grail" of implementing CMOS PAs is not that in and of itself, but the ability to integrate CMOS PAs with entire transmit and receive chains, and to generate a single-chip radio. In order to accomplish that, the PA must co-exist on the same substrate as all the other transmit and receive functional blocks. Especially in the case of PAs which generate large amounts of output power and large signals, large signals will be sent into the substrate on which the PA sits. Many state-of-the-art CMOS processes use wafers which have a substrate layer with extremely low resistivity. This substrate will easily conduct signals across great chip distances, potentially corrupting other blocks on the chip. In the case of the integrated transceiver, the goal will be to integrate extremely sensitive analog circuits such as the frequency synthesizers with the PA and even potentially digital circuitry. The PA signals can corrupt the frequency synthesizers, as they are already susceptible to a phenomenon known as LO pulling, in which frequency synthesizers react to signals injected at slightly different frequencies or with slightly different phases and lock to those injected signals rather than the crystal frequency. This distortion can severely impair the operation of the overall transmitter. In addition, if the PA were to be used in a full-duplex transceiver (i.e. one in which the transmitter and receiver must operate at the same times), extreme care must be taken such that the PA signals do not degrade the ability of the receiver to sense the extremely small signals that they must sense. In other words, the PA should not degrade the sensitivity of the receiver. In the end, precautions must be taken that the large substrate currents and voltages that the PA can generate do not corrupt other portions of the circuit.

Moreover, the substrate layer has a detrimental effect on spiral inductors built above it. Because the substrate is of a low resistivity, it allows for the creation of eddy currents due to the magnetic field generated by the planar inductor structure. These eddy currents reduce the effective magnetic field, which in turn dramatically reduce the quality factor (Q) of the spiral inductors[30]. This reduction in the Q of the spiral inductors is an extremely critical issue, as the spiral inductors are key to generating gain at the RF frequencies of interest in the design of wireless transceivers. As stated in Section 4.5, the use of inductors is critical in RF applications, as the impact of the capacitances present in the circuit must be mitigated in order to generate the required gain. Generally speaking, the Qs of inductors available in processes that use higher-resistivity substrates (such as silicon bipolar and GaAs) are significantly higher than the Qs available in CMOS processes. In fact those Qs can be on the order of 10 or more[30], while the Qs of inductors in CMOS processes are limited to 5 or less. This is yet another reason why these other technologies often have better performance for these RF circuit blocks and applications than CMOS does.

4.7 Solving the Technology Issue

The limitations associated with implementing PAs in CMOS listed in this chapter may seem overwhelming. However, there are methods that can be used to counter some of the limitations that exist. The challenge of the circuit designer is always to find ways around the problems that either technology or physics presents him; this case is no different. Chapter 5 will detail some solutions to the problems listed here, and Chapter 6 will describe the design and implementation of a prototype that will put all the work done here to the test.

5 Circuit Design Techniques

5.1 Introduction

In the previous chapters, both the issues of the design methodology and the limitations on PA implementations inherent in CMOS were discussed. Once these issues have been understood, methods of overcoming some of the limitations introduced by the use of a CMOS process must be found. In practice, the PA should have a reasonable efficiency, and if one were to design a PA without trying to minimize the impact of the limitations discussed earlier, the efficiency would in all likelihood be extremely poor. Generally speaking, the more current tat needs to be driven to the output (required to counter the oxide breakdown problems discussed in Section 4.2), the lower the efficiency will be. As stated earlier, the increased current requirements cause more power to be dissipated in the parasitic resistances inherent in the circuit and require larger signal drives. Both of these issues cause the overall efficiency to drop, and the reduction in efficiency can be quite severe in a standard CMOS process. Therefore, circuit design techniques which mitigate the impact of these limitations must be identified. In this chapter, several of these techniques will be discussed in greater detail.

5.2 Circuit Solutions

In Chapter 4, several problems inherent to CMOS processes were introduced, including oxide breakdown, poor transconductance, large device sizes, and substrate coupling. Methods of skirting these issues will be presented in this section. While there may not be a one-to-one correspondence between a particular method and the potential problem it addresses, all of the methods listed will in some way attempt to solve some of these problems and improve the overall performance of a CMOS PA.

5.2.1 Differential Structure

The first method of addressing the problems presented in Chapter 4 is one that is common to circuit designers. Differential topologies are extremely useful in that they are ideally immune to common-mode signals and prevent any noise that might exist on the power-supply from impacting the circuit performance. However, the standard well-known differential structure as used in operational amplifiers (opamps) needs to be adapted to be used in a high-efficiency PA.

The differential pair, the simplest of differential circuits, is shown in Figure 5.1. The structure consists of two transistors whose sources are connected and feed a tail current source. The drain of each transistor is loaded with an equal load impedance. The transistors are driven with a signal which consists of a common-mode signal (V_{CM}) and a differential-mode signal (v_{id}). Through superposition, the gain to the output due to the common-mode input and the differential-mode input can be separated and it can be proven that the differential-mode gain is significantly larger than the common-mode gain. One of the elements of the differential pair that contributes to the common-mode

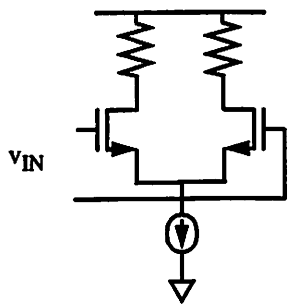


Figure 5.1. Differential pair

rejection is the fixed tail current source. That fixed current source forces another constraint on the operation of the circuit. Due to the tail current source, the node common to the sources of the two devices acts like a virtual ground for differential signals and a virtual open-circuit for common-mode circuits. However, this benefit is tempered by the fact that the tail current source also places a limit on the gain that can be achieved in the differential pair. As the magnitude of the differential input swing increases, more of the tail current will pass through one side of the differential pair and then the other. This increases until all of the current passes through one side of the differential pair and then back to the other side. But a further increase in the amplitude will not force more current into either of the devices, as the tail current source prevents any larger amounts of current from passing through the differential pair as a whole. This then puts a limit on the gain achievable in the differential pair, which will be an issue in PA design, as discussed later on in this section.

The differential structure is quite important in the design of integrated radios. In such an implementation, all the individual functional blocks sit on the same silicon substrate. As stated in Section 4.6, signals injected into the substrate by one block can travel through a low-resistivity substrate and reach another block elsewhere on the chip. These signals will show up as a common-mode signal on the substrate terminal of the devices in the second block; in order to ensure that the impact of those signals is reduced, the implementation of all blocks, and especially those that deal with small-amplitude signals or those that are extremely sensitive, should be differential. The use of differential circuits for rejecting these substrate coupling effects is especially critical when implementing a frequency synthesizer (or more than one) on the same substrate as the PA; the frequency synthesizers are extremely sensitive and are subject to LO pulling, as stated earlier in this work. Thus the entire signal path of the transmitter should ideally be implemented using differential circuits.

In the case of the power amplifier, there is also a benefit to using a nominally differential implementation, but the reasons are significantly different than for the other blocks in the transmitter chain. Where the rejection of common-mode signals is the primary reason for using a differential topology in blocks like the frequency synthesizers, in the PA, the signals of interest are significantly large that injection of noise from other blocks in the circuit is of less importance. The two primary reasons for moving to a differential implementation in the PA are voltage swing and the frequency of substrate injection. These two points will now be discussed in greater detail.

The most significant benefit that the differential implementation provides in the design of CMOS PAs is the doubling of the available voltage swing. In Section 4.2, the issue of oxide voltage breakdown and the resulting limit on the available voltage swing was discussed. By using a differential implementation, that voltage limitation now applies to each side of the differential circuit, effectively providing twice the available voltage. This can be seen visually in Figure 5.2. Figure 5.2(a) shows the case

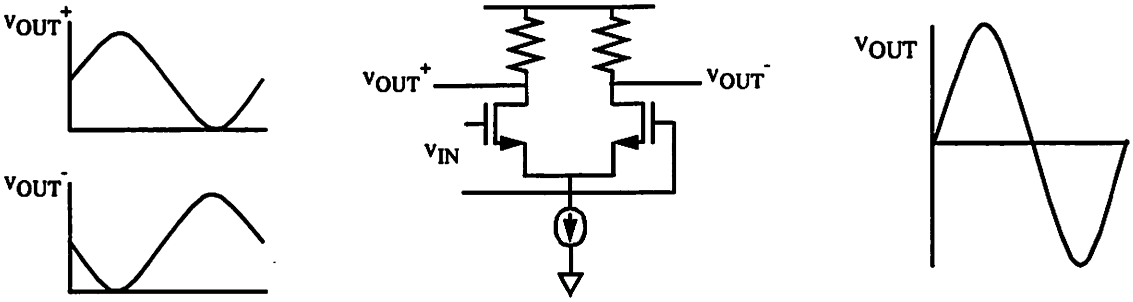


Figure 5.2. Doubled Voltage Swing in Differential Topology

of a single-ended implementation and the available voltage swing, whereas Figure 5.2(b) shows the case of the differential implementation and the available voltage. In the differential case, the output is placed across the differential output terminals. With each side of the output being able to swing as much as the single-ended case, the output voltage (i.e. the voltage across the load) is effectively doubled. However, it should be realized that this is *not* equivalent to using a single-ended implementation in which the available voltage is doubled. In that case, the amount of current required would be reduced by a factor of two. In the case of the differential implementation, the current consumption is unchanged from the original single-ended implementation; each side consumes half the original current. In essence, each side is providing half the power of the original single-ended implementation: the maximum voltage swing and half the current. Summing the powers of each side delivers the full desired power to the load.

The standard differential pair, using the tail current source, is not an ideal differential structure, however. The standard differential pair has a tail current source which limits the maximum current and thus gain available. In order to get the desired output swing and enjoy the full noise immunity benefits of the differential pair, the tail current source must sink the full amount of current required in each leg. For example, if the peak amount of current that must be delivered to the load is 500 mA, the tail current source must sink 500 mA of current. In that case, the PA will be consuming a constant 500 mA of current, dramatically reducing the PA's efficiency. Moreover, draining that much current constantly fulfills the requirement of the Class A PA, not the Class C architecture that is to be implemented in this work.

One method of reducing the current drain is to remove the tail current source altogether, and tie the differential devices directly to ground, as shown in Figure 5.3. In this topology, the DC current can be set independent of the required maximum current. The DC current is set with the bias point of the input devices, and the maximum current that the circuit can generate is set by the voltage swing at the input. Ideally, assuming

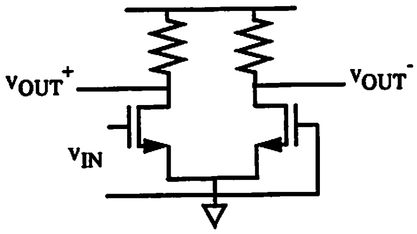


Figure 5.3. Source-grounded Differential Topology

the device stays in saturation, the maximum current is unlimited, as the current will keep increasing as the voltage swing increases.

As stated earlier, the primary benefit of this architecture is that the effective voltage swing available is increased, reducing the amount of current required to generate the same output power. This reduction in current allows the devices used in the output stage of the PA to be more reasonably sized, as each device is now approximately one-half the size of the device needed for the single-ended implementation. This is due to the fact that while the total current drawn from the supply will remain the same, the amount of current drawn by each side will be reduced by a factor of two, as stated earlier. While this does not increase the efficiency, the overall design of the PA becomes more reasonable, as the output stage devices become more easy to drive with the required swing.

Furthermore, a secondary benefit is achieved through the use of the differential topology, one that becomes extremely important when integrating the PA on the same substrate as other sensitive transmitter circuits. The large devices used in the signal path of the PA will have significant amounts of capacitance between the signal path nodes and the substrate. The consistent voltage fluctuations across these capacitances will inject significant amounts of current into the substrate, leading to the problem of substrate coupling throughout the transmitter, especially with respect to the LO-pulling problem. However, in the differential case, current will be injected into the substrate twice every cycle (one time for each side of the differential circuit), not once every cycle as in the single-ended case. As a result the frequency of the signal injected into the substrate will be twice the RF carrier frequency. Since this frequency is significantly far from the frequency of operation of the LOs, the potential for LO-pulling is reduced further.

The approach of using differential circuits provides two significant advantages over the standard single-ended case. However, further circuit techniques are still necessary in order to generate a Class-C PA design that is feasible to implement.

5.2.2 Cascode Structure

Another well-known analog circuit technique is the use of the cascode circuit structure. The cascode is generally used in operational amplifiers and other low-frequency analog designs for two reasons. First, the cascode provides an enhancement of the small-signal output resistance over the single-transistor gain stage. Second, the cascode also reduces the impact of the so-called Miller capacitor, by reducing the gain across the feedback capacitor of the MOS device. The cascode structure, equivalent to a common-source stage feeding a common-base stage, is shown in Figure 5.4. As shown in the figure the gate of the upper transistor is biased to a constant DC voltage, and the input signal is driven into the gate of the lower device.

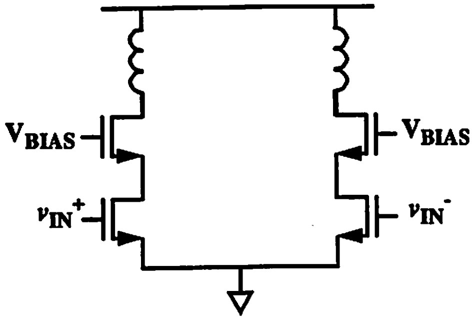


Figure 5.4. Cascode Structure

Again, though, in the case of the RF power amplifier (PA), a commonly-used analog circuit block has a different benefit. The cascode structure insulates the output node from the input node. In other words, there is no direct connection between the output node and the input node. This is extremely beneficial in the design of a PA, as the impact of oxide breakdown is greatly reduced. If the bias of the gate of the cascode device is set appropriately, the maximum stress on the oxide of the cascode device is

$$V_{OX(MAX)} = V_{OUT(MAX)} - V_{BIAS}, \qquad \text{Eq. 5.1}$$

where V_{BIAS} is the bias voltage on the gate of the cascode (upper) device. In the case of the single device stage, the maximum oxide stress is

$$V_{OX(MAX)} = V_{OUT(MAX)} - V_{IN(MIN)}, \qquad \text{Eq. 5.2}$$

which places a severe limit on the available output voltage swing. In the case of the cascoded structure, the oxide stress on the lower device is now limited to

$$V_{OX} = V_{CASC} - V_{IN}, \qquad \text{Eq. 5.3}$$

which may or may not be a problem, depending on the voltage excursion of the cascode node voltage. As stated earlier, the maximum stress on the oxide occurs when the input voltage is at its minimum, and the drain voltage of the device is at its maximum. In the single-device case, the choke inductor connected to the output charges up the drain-bulk capacitance when the device is off, raising the output voltage above the supply voltage, and increasing the stress across the oxide. In the case of the cascode implementation, the cascode node voltage will remain relatively fixed when the input signal is at its minimum, and that is because no current flows through the devices, and once the upper device shuts off, there is no current with which to further charge the drain capacitance of the lower device. Therefore, to first order, the maximum voltage on the cascode node will be limited to

$$V_{CASC(MAX)} = V_{BIAS} - V_T. \qquad \text{Eq. 5.4}$$

The maximum voltage stress across the oxide of the lower device is thus set by

$$V_{OX(MAX)} = V_{BIAS} - V_T - V_{IN(MIN)}, \qquad \text{Eq. 5.5}$$

which is a much more reasonable limit than in the single-ended case. The maximum output voltage is now increased to

$$V_{OUT(MAX)} = V_{BIAS} + V_{OX(MAX)}, \qquad \text{Eq. 5.6}$$

where $V_{OX(MAX)}$ is the maximum voltage the oxide can sustain without damaging the oxide, as explained in Section 4.2. It is apparent that the maximizing the bias voltage of the gate node of the cascode device will allow for the largest possible output swing, reducing the amount of current that needs to be drawn from the supply in order to deliver required output power.

However, the cascode does come with a few disadvantages. First, when the cascode structure is conducting current, especially in the case of a PA, the upper device will act as a switch (so that the output swing will be maximized). The on-resistance of the cascode device will dissipate power, due to the current flowing through it, incrementally reducing the efficiency. If the cascode device is large enough, however, the on-resistance can be quite small, and the resulting efficiency reduction is minor compared with the potential increase in efficiency due to the use of the cascode structure. A second issue with the cascode structure is that the current I_{DS} is reduced if a cascode device is used. In other words, if all other terminal voltages are kept the same, the output current of the cascode structure is reduced over the single device implementation. In order to compensate, the input drive should nominally be increased; again, however, the penalty associated with using the cascode is relatively minor when compared to the overall benefits.

In summary, the use of the cascode structure increases the amount of voltage swing available at the output by shielding the output from the input. The MOS gate oxide no longer sees the full difference between the input and output voltages across it, and thus the output voltage can swing higher than in the case of the single-device implementation. Therefore, less current needs to be drawn from the supply in order to generate the required output power, which is extremely beneficial. Furthermore, the sizes of the output devices required will be smaller, reducing the amount of drive that the pre-amplification stages must provide in order to deliver the required output power.

5.2.3 Inductive Tuning using Bond Wires

It has been stated throughout this work that the use of inductors in the design of CMOS RF circuits is extremely important. The use of inductors ameliorates the deleterious effects of the large capacitances seen in RF circuit blocks, especially when dealing with PAs. Because the impedance of a capacitor decreases with increasing frequency, it becomes more and more difficult to generate large voltage swings at RF in a CMOS process. In this case, "more difficult" is a simplified way of saying that more current is required to generate the same voltage swings. The use of inductors is particularly important in this work because higher current levels are already required due to many of the CMOS limitations discussed in Chapter 4. Further requiring more current would be a severe detriment to the performance of a CMOS PA. Therefore, the use of inductors is critical in this work.

However, the standard spiral inductors that are available on-chip are quite poor. The quality factor (Q) of the on-chip inductors available in standard CMOS processes is quite low. This causes problems for two reasons. First, the lower Q of the inductors presents a lower overall impedance for the tank Q. This is best explained if the situation is viewed analytically. Assume a simple LC tank has an ideal capacitor (infinite Q) in parallel with an inductor with a finite Q. Assuming the simplest definition of Q, in which the Q is determined by the ratio of the imaginary part of the impedance to the real part of the impedance, i.e.

$$Q = \frac{\omega L}{R}, \qquad \text{Eq. 5.7}$$

the parallel tank can be represented as shown in Figure 5.5(a). The previous assumption means that $\omega L/Q$ can be used as the resistor value (rearranging Eq. 5.7).

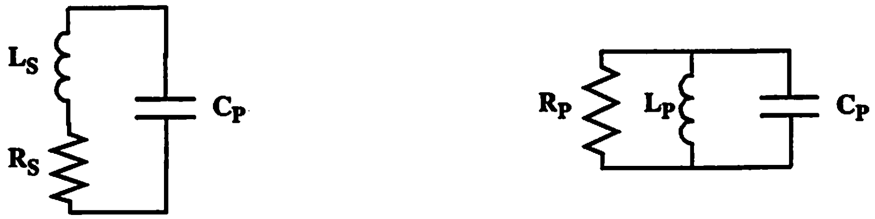


Figure 5.5. Parallel LC Tank with finite Q Inductor

The actual impedance of that parallel tank can then be determined through some simple circuit analysis, and can be determined in terms of the Q of the inductor. If the finite-Q inductor is examined, a the series R_S-L_S circuit can be replaced by an equivalent parallel circuit consisting of R_P-L_P, as shown in Figure 5.5(b). The values of the equivalent parallel inductor and resistor R_S and L_S are

$$R_P = R_S(Q^2+1) \approx Q^2 R_S \qquad \text{Eq. 5.8}$$

$$L_P = L_S\left(\frac{Q^2+1}{Q^2}\right) \approx L_S. \qquad \text{Eq. 5.9}$$

With these equations, the tank shown in Figure 5.5(a) can be replaced by an equivalent tank as shown in Figure 5.5(c). The equivalent impedance of the tank, therefore is approximately Q^2R, which increases linearly as the Q increases for a given inductance value. The parallel inductor and capacitor, which are both ideal now that the finite resistance of the inductor has been separated out, will resonate. The remaining resistor provides the finite impedance of the overall circuit. A key factor in the ability of the PA to provide gain, therefore, is the available Q of the inductors used in the circuit.

Another problem that is related to the Q of the inductors used in the circuit is the issue of power dissipation. As the series resistance of an inductor increases (and thus its Q decreases), the amount of power dissipated in that resistance itself increases as well. This power dissipation reduces the power delivered to the load and thus the efficiency. This is especially important at the output, where the optimum output resistance will be quite small (due to the issues of limited voltage swing explained earlier). If the Q is low, that series resistance can create a significant voltage division when placed in the circuit with the load resistance. This may be more clearly seen in a graphical example. Using the step-down network first shown in Figure 3-5, the output of the circuit with the finite Q inductors will look something like that shown in Figure 5.6. If it is assumed that the assumed that the L-C step down network converts the real load impedance of 50Ω to the desired optimum output resistance, then the series resistance of the finite-Q inductors is directly in series with that optimum resistance, and burning a good portion of that power that was intended to be delivered to the load. Furthermore, if the choke inductor L_C is supposed to resonate with the drain capacitance and provide a bandpass filter that allows the RF carrier to be passed to the output but rejects the higher order harmonics, the Q of those choke inductors is critical. A reduction in the impedance presented by that LC tank will increase the amount of the signal at the RF carrier frequency that will be shunted into the "filter" as opposed to the output. This further reduces the output power available delivered to the load for a given amount of output current generated by the final stage of the circuit.

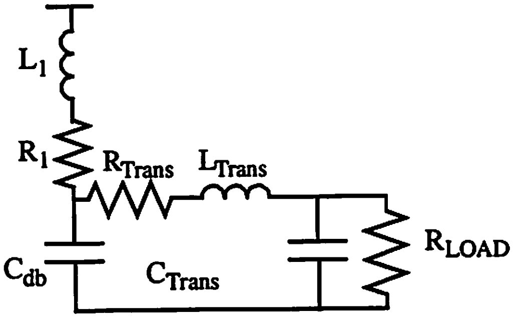


Figure 5.6. Output Network Including Finite Q Inductors

As stated in Section 4.6, the Q of an on-chip spiral inductor is generally extremely poor, normally not much better than five in a standard CMOS process[30]. To better understand what the impact of a Q of five is, an example shall be introduced. A 1nH inductor has an impedance of approximately 11Ω at 1.75GHz. With a Q of 5, the series resistance inherent to that inductor is approximately 2.2Ω. Further, if the optimum impedance that the circuit wants to see is on the order of 5Ω, it is easily apparent how much of an impact the Q has on the circuit. The existence of that 2.2Ω resistance reduces the power delivered to the output by 30%!

It is apparent that the use of on-chip inductors at the output of the final stage is not particularly feasible in a standard CMOS process, unless either the substrate is modified to reduce the impact of eddy currents (and thus increase the inductor Q) or some other method of increasing the Q is realized. Another possible way to increase the Q of certain inductors is to use inductors other than on-chip spiral inductors, which have the aforementioned Q limitations. One option is to use off-chip inductors, which can have much higher Qs and thus provide both the gain and reduced parasitic power dissipation that the inductors should have in this environment. However, one problem with this approach is that the inductor values required in the circuit can be quite small, often on the order of or significantly less than the parasitic inductance of the bond wires connecting the chip to the package that it sits in.

This in itself results from two problems: first that the package parasitics can force the bondwire parasitic inductances to be excessively large, and second that the bondwire parasitic inductances would impact the output network. These problems can be turned to the advantage of the circuit, however, if two simple steps are taken. First, a packaging technique in which the parasitic value of the bondwire inductors is reduced must be used. Second, a method of obtaining a higher Q inductor (be it off-chip or some other method) must be discovered as well.

A simple packaging technology that reduces the parasitic inductance of the bondwires used is chip-on-board (COB) technology. In this technology, the die is directly bonded to the test or product Printed Circuit Board (PCB). Landing zones around the die-attach area on the PCB are used to bond from the pads on the die to the board itself. The separation on the board between the die-attach area and the landing zone for each individual wire is limited only by the manufacturability of the board design. It is not uncommon to have that minimum spacing be as small as 3 mils (where

1 mil = 1/1000 inch ≈ 25.4μm). Thus the minimum distance between the die and the landing zone can be as small as 3mils or smaller. While the inductance of the bondwire is not an extremely simple function of its length, a good rule of thumb is that the inductance is approximately 1nH per millimeter in length. Therefore, as the minimum distance decreases, the minimum bondwire length also decreases; this can reduce the bondwire inductance to a value which can be accounted for much more easily in these designs. An example of the COB packaging method is shown in Figure 5.7[6].

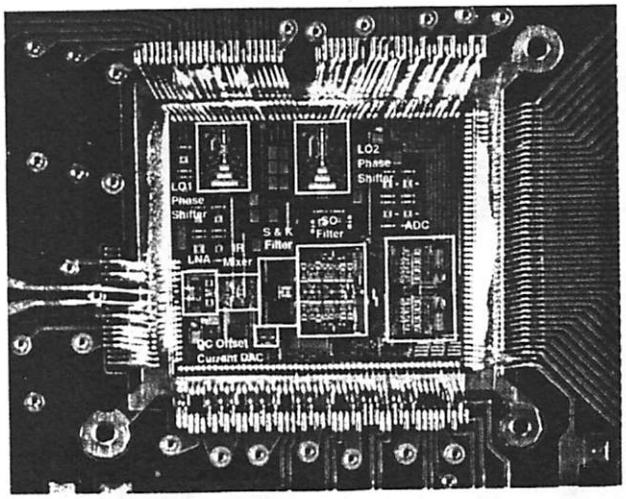


Figure 5.7. An Example of Chip-on-Board (COB) Packaging

An added benefit of using this packaging technique is that it helps to reduce the inductance of the bondwires which are used to make the supply and ground connections. In a standard package, the inductance of the bondwires can be on the order of a few nanohenries; at 1.75 GHz, that can be an impedance of as much as 50Ω or more. For standard low-current analog blocks, where the AC current might be on the order of a few hundred microamperes, at its peak, that 50Ω impedance is negligible. However, in the power amplifier, the desired amount of current flowing through the bondwire is on the order of hundreds of milliamperes or even one ampere at its peak. In this case, as the current ramps up, the voltage across the ground and supply bondwires will increase as well, reducing the effective supply range and the VGS that each device sees, reducing the overall current. With impedances in the tens of ohms, it seems virtually impossible to drive the hundreds of milliamps necessary to generate the required output power. Using even a 1nH impedance for the bondwire means that the bondwires to ground need to support an AC signal larger than 10V! It is apparent that the bondwires for the PA must be extremely short. The COB technology is another way to reduce the impact of this problem as well, as the minimum length of a ground wire can be extremely small. If the PCB layout is designed to look like the layout in Figure 5.8, the distance of a ground bondwire from the pad on the die to the ground connection on the board can be quite small. The die-attach area will normally be electrically connected to ground. A slight extension of the die attach area in those regions where the pads for the ground bondwires are allows the bondwires to be bonded directly to the die attach area. This distance can be very short, significantly less than 1mm, minimizing the inductance of the ground bond wires. The ground inductance can further be reduced by using multiple ground bondwires, although the effect of the parallel bondwires will be limited by the mutual inductance of the bondwires. In this way, the ground inductance can be made much more reasonable.

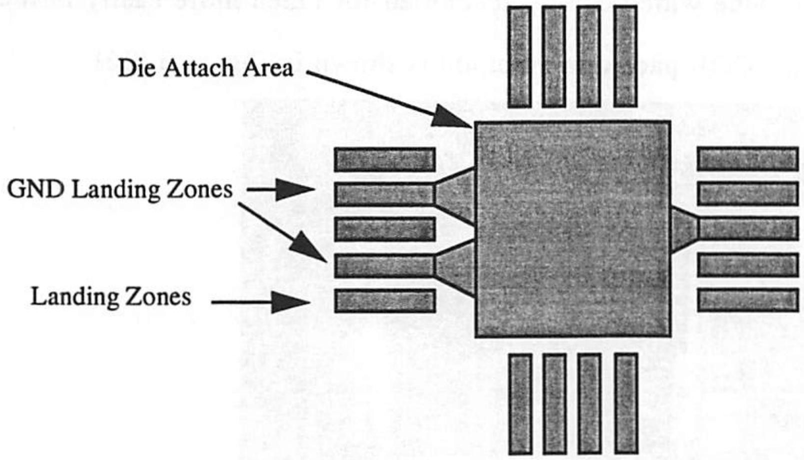


Figure 5.8. Chip-on-Board Die Attach Layout

However, now that this first step (the COB packaging) has been taken, a second step must be taken in order to have high-Q, low-inductance elements available. That second step is to use the bondwires themselves as the high-Q inductors. Depending on the wire material used, bondwires can have Qs ranging between twenty and fifty, significantly more than the peak Qs of four or five available from an on-chip spiral inductor. Standard aluminum wire, the most common bondwire material, has a Q of about twenty to thirty, which should be quite adequate for the application in this work. The primary disadvantage of using just a bondwire as an inductor in a circuit is that the bondwire inductance is difficult to control. When actually placing a bond wire, many parameters can affect the final inductance value of the bond wire. The horizontal length, the height of the loop, and the angle of the wire are among a few of the determinants of the inductance of the bondwire. Furthermore, the overall inductance value of a bondwire is also impacted by the orientation and layout of bondwires adjacent to it, as the mutual inductance between one bondwire and its many neighbors will affect the inductance. Therefore, while an approximate idea of what the bondwire should look like physically may be known, actually implementing that can be somewhat difficult, and usually requires a process of iteration to implement a bondwire of the correct value. While the initial efforts to implement a bondwire with a particular inductance may prove difficult, once the geometry for a given inductance value is known, it is not difficult to duplicate that geometry, especially when using an automated bonding machine[33]. Thus if the geometry can be determined, the implementation can easily be expanded to a production line requiring large-scale repetitions of the bondwires. A secondary disadvantage of using bondwires as inductors is that each bondwire will generally need its own pad, which may be difficult as many analog designs encounter a dearth of space for pads. It may therefore be difficult to use bondwire inductors for each and every inductor required in the circuit.

The use of bondwire inductors and the Chip-on-Board (COB) packaging technique can provide for high-Q inductors, even in an integrated CMOS technology. These techniques have some drawbacks, but the implementation of PAs in CMOS without these techniques is extremely difficult.

5.2.4 Input tuning of final stage

Even with all these methods of facilitating the design of a PA for cellular standards in CMOS, the implementation can still be difficult. The size of the output devices that may be required in the design can still be excessively large, as the current needed to generate the required power, while less than it would be if the above techniques were not employed, can still be excessive. As stated earlier, the problem with the output devices being too large is that the pre-amplification stages will have difficulty in actually driving the output stage. The pre-amplification stages will have to sink larger amounts of current to generate the required signal drive, further reducing the efficiency of the overall PA implementation. If a high-impedance node at the input of the output devices existed, the task would be simplified, as less current would be required to generate the same voltage swing at the input of the transistors. However, with exceedingly large devices, the gate-capacitance will present a very heavy load to the pre-amplification stages.

Certainly, inductors can be used at the output of the previous stage in order to create that desired high-impedance node. However, two problems are apparent. In order to understand these problems, a simplified AC model of the nodes between the output of the preamplification stages and the input of the final stage is shown in Figure 5.9. The coupling capacitor between the two stages is required so that the input to the final stage may be DC biased separately from the output of the previous stage, as is

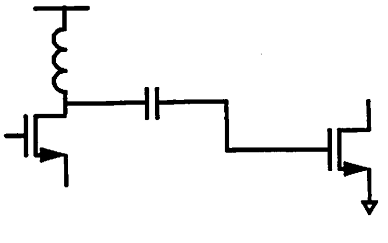


Figure 5.9. Interstage Matching at Input of Output Stage

required in a Class-C implementation. In an integrated environment, the implementation of a interstage capacitor will also have a significant amount of ***bottom-plate capacitance***, which is the parasitic capacitance to the substrate from the bottom-plate of the capacitor implementation. In the case of a double-poly process, this bottom-place capacitance may only be 20% of the value of the desired capacitor. However, in order to prevent too much signal loss, the coupling capacitor should be large enough that the voltage division across the capacitive divider is reasonable. That is, in order to limit the loss to about 20%, the size of the coupling capacitor can be estimated to be

$$\frac{C_C}{C_G + 0.2C_C + C_C} = 0.8 \Rightarrow C_C = 25C_G. \qquad \text{Eq. 5.10}$$

The coupling capacitor has to be twenty-five times the gate capacitance of the final stage in order to pass 80% of the signal! Even if the amount of signal that is passed across the coupling capacitor is reduced to 70% of the original signal, the required value of the coupling capacitor is still more than six times the value of the gate capacitor. While, in general, area is somewhat readily available in a CMOS process, Eq. 5.10 indicates that the value of the coupling capacitor must be on the order of several hundred picofarads, which is difficult to integrate, even in a CMOS environment. Furthermore, the amount of capacitance that the previous stage must drive is 4.8*C_C, which can still be on the order of a hundred picofarads. In order to tune out a hundred picofarad load, an inductor of 0.08nH must be implemented. This inductor is not practical to implement, and therefore a different method of tuning must be found.

The previous inductive tuning method attempts to create a high-impedance node at the output of the previous stage, and then to allow for losses between that node and the input to the final stage. If, instead, a high-impedance node were created not only at that node but at the input of the final stage, a more practical implementation might be achieved. However, simply placing an inductor in parallel with the gate capacitance of the final stage will not be feasible, as the DC bias will then be set by the inductive path. Ideally, the inductor should tie to a DC source or sink rather than a voltage generation circuit, as the amount of current passing through the inductor may be significant. A simple method of creating this high-impedance node is to use a series inductor-capacitor circuit in parallel with the gate capacitance of the output, as shown in Figure 5.10. This series L-C serves two purposes: first, it eliminates any DC path through the inductor that might set the bias of the gate of the output stage, allowing the bias to be set independently, and second, the equivalent capacitance that the inductor sees is reduced, allowing for a larger (more practical) inductor to be used. Working through the math, the equivalent capacitance that the inductor sees can be represented as

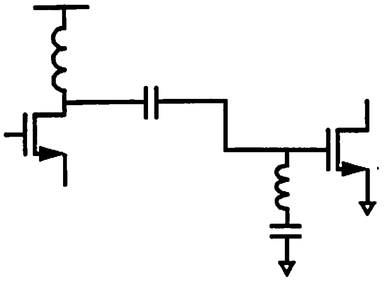


Figure 5.10. Modified Tuning Method

$$C_{EQ} = \frac{(C_G + 0.2C_C)C_S}{(C_G + 0.2C_C) + C_S}. \qquad \text{Eq. 5.11}$$

The capacitor seen by the inductor is the series equivalent of the two capacitors in the circuit, as shown in Figure 5.11. The term 0.2*CC represents the parasitic bottom plate

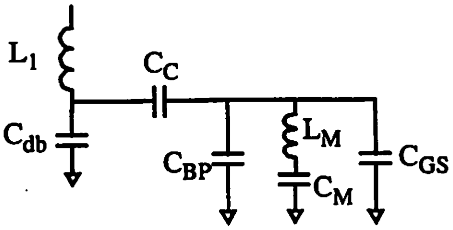


Figure 5.11. Equivalent Circuit for Input Node of Output Devices

capacitance of the coupling capacitor, as discussed earlier. The parallel combination of the series L-C and the gate capacitance C_G creates a high-impedance node at the input of the output stage. That high impedance node reduces the value of the coupling capacitor needed to minimize the loss and preserve a large signal.

As an example, assume the gate capacitance of the final stage was 30pF, and the output capacitance of the previous stage was approximately 7p. The equivalent circuit for this is shown in Figure 5.12(a). In order to ensure that 80% of the signal passed to the input of the output stage, the coupling capacitor required is on the order of 188pF. The equivalent capacitance seen at the output of the second stage (where the tuning inductor will be placed) is then 57p, which requires a 0.15nH inductor in order to tune out that capacitance. Implementing a 0.15nH inductor in a CMOS process is very impractical, independent of whether a bondwire or an on-chip spiral inductor is used.

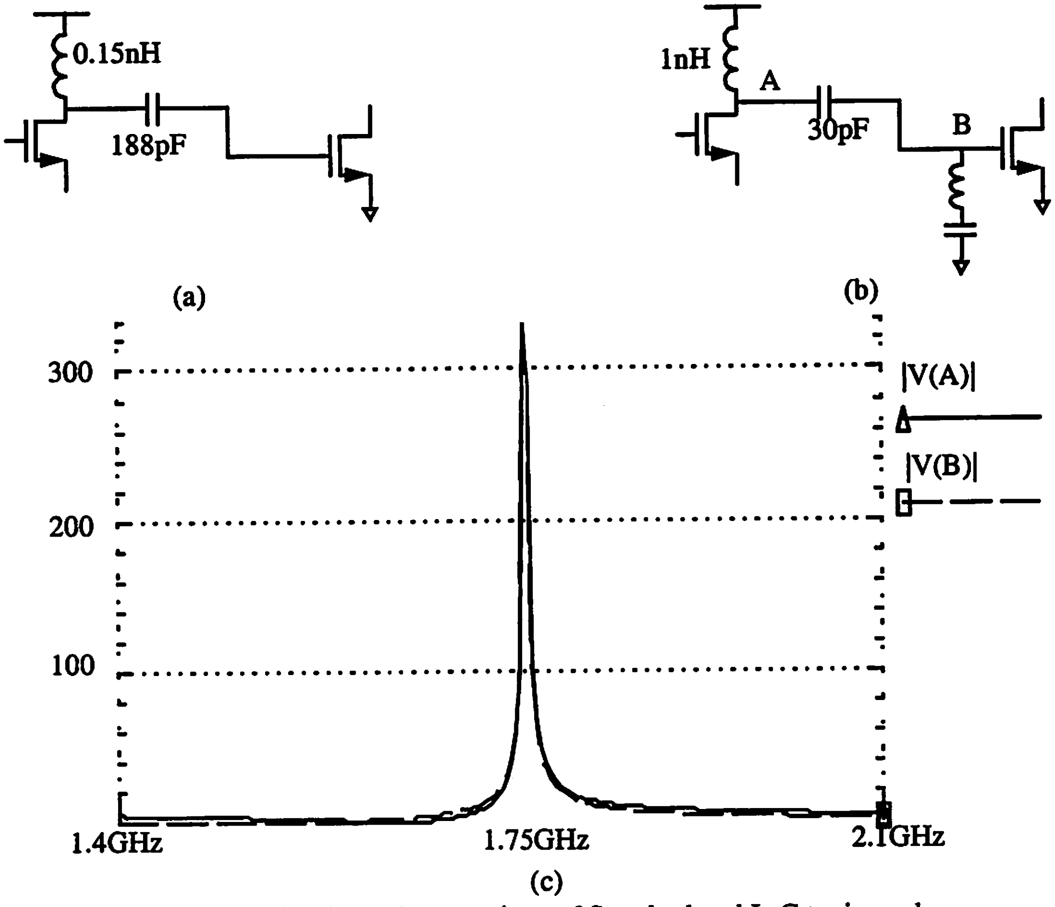


Figure 5.12. Circuit Implementations of Standard and L-C tuning schemes

Using the series L-C technique described above, however, a much more reasonably-sized coupling capacitance can be used with the circuit, as well as reasonably-sized inductors as well. Shown in Figure 5.12(b) is the implementation of the same two capacitors, this time using the series L-C network in order to create a impedance peak at the input to the final stage. The coupling capacitor is now only 35pF, and yet the signal loss across the coupling capacitor is quite small. Figure 5.12(c) shows the output of a simple AC simulation from HSPICE, in which the transfer functions at both the output of the pre-amplification stage and the input of the final stage are shown when the network in Figure 5.12(b) is driven with an AC current source. Ideally, the loss across the coupling capacitor is extremely small, thus maximizing the amount of signal which is delivered to the final stage. In Figure 5.12(c), the amplitude of the signal at both nodes A and B are approximately the same at 1.75GHz (the curves are right on top of each other); ideally, little signal is lost across the coupling capacitor. In reality, the performance is not as good, but it still allows for increased signal transmission across the coupling capacitor using smaller components.

5.2.5 Cascode Inductors

Finally, one other technique was used in this design. It has been previously stated that there are several large capacitors existing in the circuit, usually resulting from the drain-to-bulk capacitances of the large devices used. Those capacitors in the signal path must be charged and discharged as the voltage fluctuations from the signals pass through the circuit. The capacitances at the output node of each stage and the corresponding input nodes of the next stage are being tuned out with inductors. Another way to think of the "tuning" process is to consider that the inductors are storing current, which is circulating between the inductors and capacitors. That stored current is charging and discharging the capacitors, reducing the amount of current that must be drawn from the supply for that purpose. If less current is required from the supply to force a certain voltage swing on a certain node, the appearance is that the impedance is larger, which is what happens when a capacitance is tuned out using an inductor.

However, there are still large capacitances in the circuit that are not being tuned out, and are draining current from the supply in order to charge and discharge as required. These are the capacitances at the intermediate cascode nodes of the stages using that structure. The voltage excursions on these nodes are not as large as those on the outputs, as the voltage on the cascode is limited on the high side by the fact that the upper device will turn off. However, these are significant capacitances, and any method of reducing the current that goes to charge and discharge these nodes will boost the overall efficiency of the PA.

To that end, a simple method for reducing that current was used. An inductor was used at the cascode node in order to "tune out" those capacitances. In other words, an inductor was used to store some of the current required to charge and discharge those nodes. However, because the inductor is a short-circuit in a DC sense, that inductor must be tied to a node with the proper DC bias. Furthermore, that node should ideally be a node that is able to provide a large amount of current, as explained earlier in Section 5.2.4. However, that would essentially require another supply voltage, at approximately the bias of the cascode node, which is difficult. Moreover, the introduction of this extra voltage supply runs counter to the reasons for the CMOS implementations. The requirement that another supply is required only increases the size and cost of the final implementation.

Another solution is available, which does not require any other discrete or external components. Because this PA is implemented in a differential fashion, a second cascode node, with the exact same bias, exists on the complementary side of the circuit. The cascode inductor can now be tied across the differential implementation. Using an inductor of twice the size of the inductor in the single-ended case, a differential implementation is quite easy to effect. The inductor would be placed in the circuit as shown in Figure 5.13. While the inductor is shown as two separate inductors

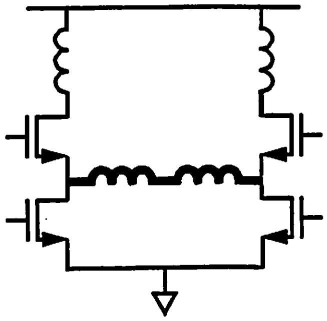


Figure 5.13. Cascode Inductor in Differential Circuit

joined together, it can certainly be implemented as a single inductor with double the value of the individual inductors 'f necessary. If this technique is used, there is a considerable boost in the overall efficiency of the PA.

One slight precaution must be taken when adding this inductor to the circuit. The voltage excursion on the cascode node was described in Section 5.2.2 in the context of the potential for oxide breakdown. There it was determined that because the upper limit on the cascode node voltage is V_{BIAS}-V_T, where V_{BIAS} is the bias voltage on the gate of the cascode device, and V_T is the device threshold voltage. However, once the cascode inductor is added, that upper bound is no longer valid. The inductor will store current, and drive current into the capacitor even when the cascode devices have turned off. Therefore, the voltage on the cascode nodes can rise above the previously set limit, and the issue of oxide breakdown must be considered again. When proceeding with the actual design, care must be taken that the rules described in Section 4.2 regarding oxide breakdown are not violated.

5.3 Conclusion

This chapter detailed several methods through which the impact of the limitations that sub-micron CMOS technologies impose on the design of PAs could be reduced. The overall goal of this work is to show the feasibility of the design of high-performance PAs in CMOS; therefore, the problems that are caused by the choice of technology must be eliminated or at least mitigated. In order to prove the feasibility of these design techniques, the next chapter will discuss the implementation of a prototype which uses all the design methods listed in this section.

6 Power Amplifier Prototype Implementation

6.1 Introduction

The previous sections described in detail several methods by which the limitations described in Chapter 4 could be ameliorated. These methods must then be used in an actual design in order to test their feasibility. In this work, a PA which can meet the specifications of the DCS-1800 cellular system (a variant of the well-known GSM standard) was designed. The prototype PA was designed to be integrated with the entire transmit and receive paths, implementing a single-chip CMOS transceiver. The prototype was also implemented in a stand-alone test-chip configuration, for two reasons. First, it allowed the testing of the PA and the rest of the transmitter to be done in parallel, reducing the overall evaluation time. Second, the performance of the PA could also be verified without requiring that the rest of the transmitter be functional as well. This chapter will cover the actual design of the prototype, incorporating all the methods and techniques described in the previous section. First will come a discussion of the specifications of the system, followed by the prototype design and layout.

6.2 Power Amplifier Specification

In order to show the viability of Class C PA, a prototype must be designed and built. In this work, a Class C PA that satisfies the requirements for DCS 1800, a version of the GSM mobile telephony standard, was implemented. As stated earlier, the PA was to be designed in a CMOS process, in order to integrate this PA with the rest of a full transceiver chain on a single chip. In DCS 1800, the peak output power is one watt at the antenna, and must be controlled to lower levels. The GSM standard does use a constant-envelope modulation scheme, so the use of a heavily non-linear PA, like a Class-C implementation, can be investigated. That is the task undertaken in this work, in order to show the feasibility of using Class C PAs for narrowband RF applications. The transmit band in DCS-1800 is between 1.71 and 1.785GHz, with 200kHz channel spacings.

One key to the narrow channel spacings and the relatively narrow band (75 MHz at a center frequency of approximately 1.75GHz) is that narrowband tuning can be used. All of the tuning techniques discussed until now have stated that high-Q inductors are desirous, which is only true if the band over which the PA must provide gain is narrow. If the system's bandwidth is a wide band, high-Q components are difficult to use, as the resulting resonant peak will have a narrow band over which the impedance peaks. Thus the narrow-band nature of the GSM and DCS-1800 specifications is very important to the design methods discussed previously.

Furthermore, an important point to understand about the output power specification is that the 1-W peak output power refers to the power delivered to the antenna. The reason that this is an important point is that there will be one or more components between the PA and the antenna which may slightly attenuate the signal before it reaches the antenna. In the case of a single-ended PA, there will be a RF filter and/or a duplexer between the PA and the antenna. The filter is there to suppress any signal harmonics or noise far outside the channel bandwidth. In the case of a system like GSM, in which the transmit and receive bands are diverse in frequency, a duplexer will specifically attenuate the harmonics and noise that the transmitter and PA generate in the receive band. In addition, in the case of a differential PA, the differential signal will have to be converted to a single-ended signal, as almost all common antennas used in cellular handsets today are single-ended. This conversion will usually be done with a balun, which will also attenuate the signal. Therefore, at the boundary between the chip and the PCB, the PA should provide more than the peak power specified by the standard. In this work, the PA was designed to provide 1.7-W peak power, allowing for more than 40% of the signal to be lost before reaching the antenna while still delivering 1-W of power to the antenna.

6.3 Circuit Design

In this section, the actual design of each of the stages will be discussed, as well as the final circuit implementation.

6.3.1 Process

The Class-C PA described in this work was designed in an STMicroelectronics, 0.35mm, five-layer metal, doubly poly (5MDP) process. The nominal supply voltage for the process is 3.3-V, and the f_T of the process is approximately 20 GHz for a device with a gate-source voltage of 0.9V. According to process engineers with STMicroelectronics, the peak voltage that the oxide can reasonably support is about 5.5-V. Higher stresses on the oxide voltage may cause breakdown of the oxide, of the sort described in Section 4.2. The process also allowed for high-resistivity poly-silicon resistors and high-resistivity diffusion resistors as well.

6.3.2 Output Impedance Transformation Network

Before the full output stage can be designed, the output impedance transformation network must be determined. This network will present the optimum resistance to the load by transforming the real load impedance of 50-Ω to the optimum resistance. Since the optimum load resistance is limited by the available voltage swing, that optimum value will be less than the 50-Ω presented by the antenna; therefore, a step-down network must be used to transform the antenna's impedance to the optimum resistance. The step-down network was first mentioned in Chapter 3, and consists of a series inductor and a parallel capacitor, as shown in Figure 6-1(a). One beneficial property of the step-down network is that the need for a series inductor translates well to the implementation of the PA. The chip relies on bondwires to make connections to the printed-circuit board (PCB) on which it sits. Since bondwires are to be used as inductors in this design, those bondwires can be used as the inductors in the step-down network, providing a high-Q inductor at the output of the signal path.

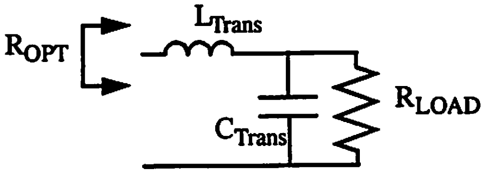


Figure 6-1. Step-Down Network

The actual optimum resistance can be determined from the known available voltage swing. The output power is related to the voltage swing by

$$P = \frac{V_{PK}^2}{2R} = \frac{V_{RMS}^2}{R}. \qquad \text{Eq. 6.1}$$

Through the use of the cascode, the peak output voltage was raised, allowing for a larger voltage swing across the load. The supply voltage is set to be 3.3V by the process and by the goal of integrating the PA with all the other blocks. If the output swings fully to twice the supply voltage and then down to 0-V, the amplitude of the swing on each side of the differential PA would be 3.3-V, and the optimum resistance is

$$R = \frac{V_{PK}^2}{2P} = 12.8\Omega \qquad \text{Eq. 6.2}$$

However, the output will not fully swing down to 0-V; as the device moves into the triode region, the two transistors between the output and the input will keep the node voltage from reaching 0-V. If it is assumed that the minimum voltage on each node is limited to 0.1-V, the approximate optimum resistance drops to 12-Ω. In the differential configuration of the output match shown in Figure 6-1(b), the values of the inductor and capacitor required are approximately 1.5-nH and 2.47-pF, respectively. Now that the optimum resistance and a first-order output impedance transformation network are known, the output stage itself must be designed.

6.3.3 Output Stage

As has been previously stated in this work, the most practical method for designing a power amplifier is to start with the output stage, and then work backwards. The critical performance specifications and the critical limitations primarily apply to the output stage and its abilities, it should be the priority. The output stage must be able to deliver 1-W of output power or more to the 50-Ω load, while not violating the process parameters listed in the previous section (especially the oxide breakdown voltage).

The basic structure of the output stage is known from the previous chapter. The output will be a pseudo-differential cascode implementation with a cascode inductor across the output; that structure was shown in the previous chapter, and is redrawn here in Figure 6-2. There are several degrees of freedom in the design of the output stage, each of which will be determined through the specifications provided by the standard and the capabilities and the limitations of the particular technology used in the design. In the design of the Class-C PA, the free variables or entities in the

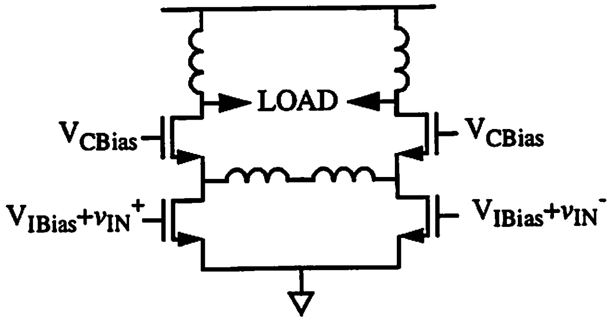


Figure 6-2. Output Stage Topology

topology are the output device sizes, the cascode device sizes, the input bias, the cascode bias, and the amplitude of the signal at the input.

The design specification calls for a 1-W peak output power at the antenna, which in this design equates to a peak output power of 1.7-W at the edge of the chip boundary. With a 12-Ω optimum output resistance, the peak current that must be driven through the load is

$$P = \frac{I^2R}{2} \Rightarrow I = \sqrt{\frac{2P}{R}} = 0.53mA\,. \qquad \text{Eq. 6.3}$$

The base devices used in the output stage must be able to drive the load with a sinusoidal current with a peak of 530-mA. Furthermore, that large a signal must be delivered from a device which is only conducting for a portion of the period, as is required by the Class-C architecture.

Of the degrees of freedom listed earlier, some of them are inter-related; that is, changing one will affect changes in one of the others. To first-order, those relations are relatively easy to identify. It should be apparent that the output device size, the input bias, and the input signal amplitude are related. As the device size increases, a lower signal drive would be required to generate the same amount of current, and the bias point may be lowered in order to reduce the portion of the period when the device is conducting. It should further be noted that the size of the cascode device may be set independently of the size of the base device. The primary purpose of the cascode device is to protect the gate of the oxide from the voltage excursions of the output node;

the base device will drive the current in and out of the output network. The cascode device is really acting as a switch when the input voltage is at its peak, and one of the limiting factors in the size of the cascode device is that the on-resistance of the device does not have a large impact on the overall efficiency of the PA. The size of the cascode device does have some impact on the amount of current that the overall cascode is able to drive, but if there are other factors which require a modification of the cascode device size, those other factors may take precedence.

Once the key performance requirements of the output stage are known, the design methodology described in Chapter 3 can be used to determine the approximate size of the output devices. The PA must be able to deliver more than 500 mA of current to a 12-Ω load, and swing up to 6-V or higher at the output. The output transformation network is known from the previous section. The simplified output circuit used in the design methodology is shown in Figure 6-3. The cascode structure is not used in the simple design methodology, as it dramatically increases the complexity of the circuit. Once the approximate first-order design has been achieved, the cascode devices can be entered and accounted for in the circuit-simulator while the design is being fine-tuned.

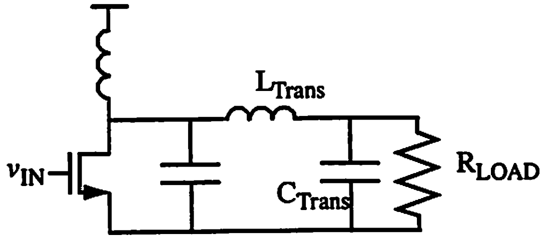


Figure 6-3. Simplified Output Circuit Schematic

It should be noted that the method to design the simple Class-C PA shown in Section 3.4.3 actually designs a PA that must deliver only one-half the required power; that is, the 125-mW PA that was described in Section 3.4.3 could actually be used to design a 250-mW PA (this was actually useful at one point; a DECT PA, which must deliver 250-mW of power, was originally to be included in this design). Since the GSM PA will be implemented in a differential fashion, the design methodology in Chapter 3 will be used to identify a "half-circuit" which will deliver one-half the required power to the load. It should be understood that the resulting implementation will perform better than the real implementation, for a few critical reasons. First, the implementation here will use a cascode, which generally produces less current for the same device size and input drive as a single-device implementation. Second, the model used in the design methodology also over-estimates the current in the device. Therefore, the actual final implementation will need to be modified to generate more current than is really provided by the implementation deriving from the design methodology.

The design methodology was used to determine that a reasonable design to deliver the required power had the following properties: a device size of 18000µm/0.35µm, input bias of 0V, and an input signal amplitude of 1.6V. From this starting point, the methods by which more current can be delivered include changing the input bias, the input amplitude, or the device size used in the PA. The device size determined above is already extremely large; any further dramatic increases would be make the device even more difficult to drive. As a result, the modifications made to the initial design were primarily in the input bias and input signal magnitude. Furthermore, one other degree of freedom is available to the designer now, and that is the size of the cascode devices to be used. In this case, there is a trade-off that must be understood. As the cascode device size is increased, the available current is increased as well; however, the output capacitance (from the drain-bulk capacitance of the cascode device) is increased also, which at some point can become difficult to tune out. As a result, simply maximizing the cascode device size may not be the optimum solution.

For this particular design, the final design of the final stage is shown in Figure 6-4. The size of the base device is slightly increased, but the input bias is increased to

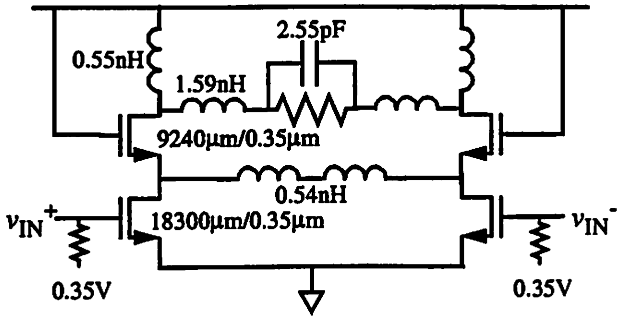


Figure 6-4. Final Output Stage Implementation

0.35-V, and the size of the cascode device is approximately half of the size of the base device. This was done in order to reduce the output capacitance, as stated earlier. The two sets of inductors at the output are both to be implemented as bond-wire inductors, in order to take advantage of the higher quality factor (Q). The cascode inductor used was implemented as an on-chip spiral inductor (the design of the spiral inductors will be discussed in greater detail later in this chapter). The use of the cascode inductor does involve one slight variation in the implementation with respect to the layout of the devices, namely that the cascode node must be accessible in order to contact with the terminals of the inductor. Standard cascode layout trades the accessibility of the cascode node with a reduction in the capacitance at that node. This will be explained in more detail in Section 6.4. In simulation, the use of the cascode inductor increased the efficiency of the output stage quite dramatically.

6.3.4 Second Stage

The second stage of the PA, which drives the output stage of the PA, is shown in Figure 6-5. The second stage of the PA was also designed as a Class-C stage, in order to reduce the power consumed in this stage. The input matching network of the output stage is also shown, including the modified tuning network described in Section 5.2.4. In this implementation, both the AC-coupling capacitors and the series L-C capacitors are implemented as on-chip poly-poly capacitors, and the series L-C inductors are implemented as bond-wire inductors. The load inductors at the output of the second stage are implemented as on-chip spiral inductors in order to reduce the pin

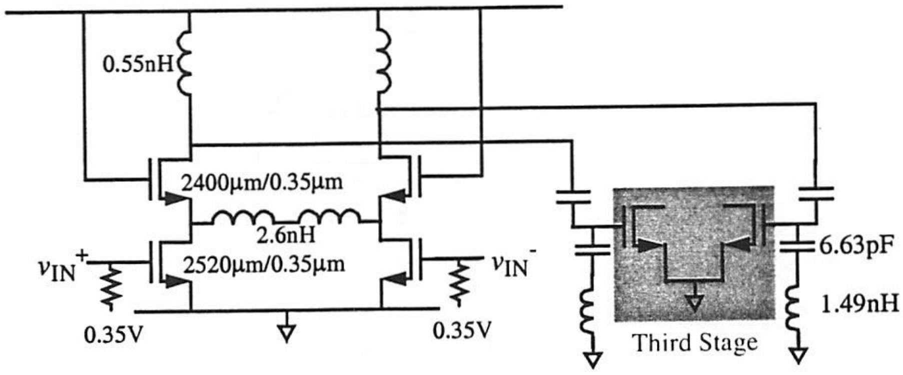


Figure 6-5. Second Stage of PA

count of the PA block. The pin count reduction is especially important for the inclusion of the PA in the single-chip transceiver, where the number of available pins is a severely limiting factor.

One other factor that helped determine the size of the second stage is the fact that a second amplification stage (referred to here as the "first stage" in the next section) is required in order to present a load that can be driven by the RF mixers (previous stage of the transmitter chain). Furthermore, in order to perform power control, a more linear (and therefore more inefficient) stage is preferable, as explained in Section 6.3.6. In order to reduce the impact of the more linear stage on the overall efficiency of the PA, another stage which both reduces the loading on the mixers and consumes considerably less power is desired. The design of the first stage of the PA is presented in the next section.

6.3.5 First Stage

The first stage of the PA in this design is the one in which the power control is performed. In order to achieve this goal, a more linear first stage was used. The first stage is simply a differential pair with a tail current source. The variation in the tail current source allows the gain throughout the PA to be controlled by the gain in the first stage. The schematic of the first stage is shown in Figure 6-6. The power control scheme is described in more detail in the next section.

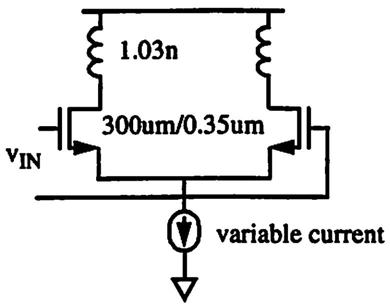


Figure 6-6. First Stage of PA

6.3.6 Power Control Scheme

The power control scheme used in this work is a relatively simple one. By varying the tail current through the differential pair shown in Figure 6-6, the differential gain through the PA is changed as well. The peak voltage swing at the output of the differential pair is related to the current available to swing through the load. In the case of the output stage and the second stage, no tail current source was used; the tail current source forces a constant DC current to be sunk in the circuit, increasing the power consumption and forcing more of a Class-A topology. That is contrary to what is required in the later stages of the PA. However, in this earlier stage, sinking a constant DC current is not as problematic as it is in later stages. The amount of power consumed in the first stage, even with constant crowbar current flowing through the differential pair, will generally be less than ten percent of the overall total.

Due to the fact that the amount of current flowing in the tail current source sets the gain, the gain will vary to first order linearly with the amount of current. This does not mean that the overall output power (or voltage, more correctly) will vary linearly with the tail current of the first stage, however; the nonlinearity of the second and third stages will create a nonlinear relationship between the input of the first stage and the output as well as between the tail current value and the output.

In order to create a modular tail current source, a fixed current was brought on-chip and then mirrored over to the actual tail current of the differential pair. The current mirror had a current gain of 10:1, so that the amount of current brought on chip was not excessive (after all, that current will affect the power consumption of the PA as well). Seven digital bits of control were provided over and above the nominal value of the tail current. The current could be varied between 0.005 and 20 times the nominal current using the digital bits. These bits were stored in a bank of shift registers, and can be loaded into the shift registers from off-chip.

6.3.7 Preamplifier

In Section 6.3.4, the need for a second stage of amplification prior to the output stage was presented. Without another stage, the loading on the mixers can become extremely large; some method of pre-amplification would be necessary. If the mixers on the integrated chip were only driving the PA, no further issues would have arisen. However, the integrated transceiver chip included two PAs as well as a testing buffer used to drive the output of the mixers off-chip (essentially bypassing the PAs so that the mixer output could be examined independent of the PAs). The combination of the three blocks to drive made it virtually impossible for the mixer as it stood to drive the required capacitive load of the three blocks in parallel. As a result, a separate preamplification buffer was required. Furthermore, a method was required by which the signal would not flow to the two paths which were not being used (of the three listed above). The circuit shown in Figure 6-7 was replicated three times, with each of the preamplification stages driving one of the three output stages (the two PAs and the

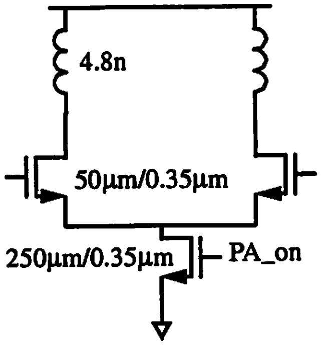


Figure 6-7. Preamplification Stage.

testing buffer). The load inductors required for each of the preamp stages is different depending on the load that is seen. The inductor values required to drive the first stage of the Class-C PA are shown in Figure 6-7. The device at the base (below the input devices) is used as a switch (with the PA_on signal) to turn the individual preamp-stages on and off.

6.3.8 Biasing

All the stages of the PA must be biased independent of the output of the stage that drives them, since each is biased well away from the supply. The stages driving each of the stages of the PA are inductively coupled to V_{DD}. In order to generate the appropriate bias voltages, a ΔV_{GS} reference current source was used to generate a fixed current. In the case of the first stage, which required a bias voltage close to 2-V, a simple diode-connected transistor was used to generate the appropriate voltage. In the case of the second and third stages, which required a bias voltage below the threshold voltage of the device, a structure similar to that shown in Figure 6-8 was used. The bias voltage was taken from the node in between the two devices.

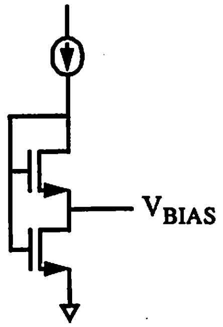


Figure 6-8. Structure used to generate sub-V_T bias voltages

In both of the bias cases, the output of the reference current source was fed into a four-bit current DAC before being driven into the bias-generating transistors. This was done to allow for the bias voltages to be modified if necessary to improve the performance of the prototype circuit.

6.3.9 Power-Down Mode

One other critical capability of the PA is that it must be able to reduce its power consumption and virtually eliminate any power delivered to the load when the transmitter is not actually on. Many cellular systems today specify that the mobile unit is only transmitting for a small fraction of the time; at other times, it is critical to reduce the power consumption of the unused blocks, in order to prolong battery life and to prevent any signals from leaking out of the mobile unit and disrupting communications.

In this PA, the way that this power-down was accomplished was to ground the inputs of all the stages of the PA, and also to turn off the tail current source of the first stage. As mentioned earlier, the first stage contained a current DAC, which could be used to control the amount of power delivered to the load. By using that current DAC to completely turn off the tail current source, as well as using switches to ground the gates of the inputs to each of the stages, the power consumption is reduced to ideally zero. The switches used to ground the inputs of the stages were controlled by bits in a shift register, which can be shifted in from off-chip. Those switches were made somewhat long, so that they would not affect the operation of the circuit in their off-state.

6.3.10 Final Circuit Topology

The final circuit topology is shown in Figure 6-9 on the next page. All the inductor and capacitor values are listed. The load inductors at the output of the preamplification stage, the first stage, and the second stage, as well as the two cascode inductors, are implemented as on-chip spiral inductors. The inductors at the output of the final stage and the inductors used at the input of the final stage in the series L-C tuning scheme are implemented as bondwire inductors in order to make use of their

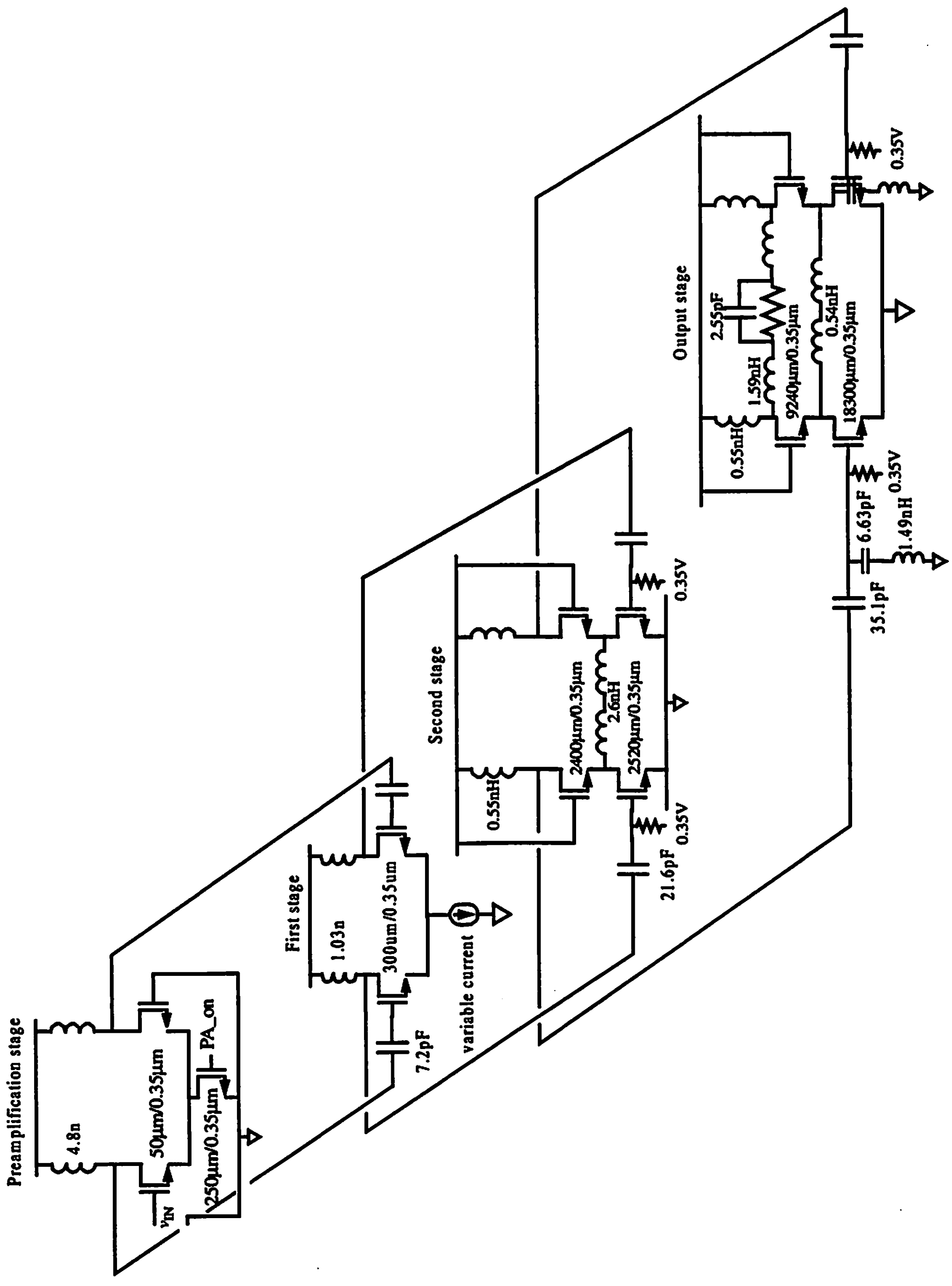


Figure 6-9. Full Circuit Schematic for Class-C PA, including preamplification stage

high-Q properties. All of the capacitors shown are implemented using poly-poly capacitors available in the STMicroelectronics process being used.

6.4 Layout and Fabrication

A critical element of the design process, especially for RF analog circuit blocks, is the layout stage. At low frequencies, the parasitic passive elements inherent in the circuit do not pose many problems; at high frequencies, however, these parasitics can dramatically impact the performance of circuit blocks and careful attention must be paid to the layout to reduce that impact. A few of the more critical elements will be discussed in this section.

6.4.1 Basic Transistor Cell

One important effect that must be accounted for is the substrate resistance [34]. At low frequencies, the drain-bulk capacitors in the MOS device present large impedances to the circuit. Therefore, even with some resistance in the bulk, the effect on the circuit is minimal. However, at high frequencies, the substrate resistance has the effect of reducing the Q of the capacitor, which reduces the Q of overall L-C tank, since the Q of two elements in parallel is the parallel combination of the Qs, i.e.

$$Q_{eff} = Q_1 \| Q_2 = \frac{Q_1 Q_2}{Q_1 + Q_2}. \qquad \text{Eq. 6.4}$$

It has previously been noted that the inductor Qs are extremely important; any further reduction in the Q of the L-C tanks used throughout the circuit can reduce the gain through the signal path and require more current to be sunk to achieve the required output power. A second problem with the MOS transistor that arises from the layout is the physical R-C due to the polysilicon gate. At minimum length, the polysilicon gate is a long, thin finger whose resistance can get large as the device width is increased. Since the polysilicon gate will be driven from the end, the series resistance (in series with the gate capacitance of the device) can introduce a low-pass filter into the circuit and reduce the signal amplitude that drives the gate.

In order to reduce the resistance between the drain/source terminals and the substrate terminals as well as the gate resistance, a base transistor cell was created. This base cell was surrounded by substrate contacts, in order to present a short path through the bulk before being collected by the substrate terminal. The base cell consists of two 10μm/0.35μm transistors with a shared drain terminal, and the gates were driven from both sides, reducing the effective width of the device with respect to the gate resistance. By using the required number of base cells required to generate the device sizes, the effective substrate resistance is further divided down.

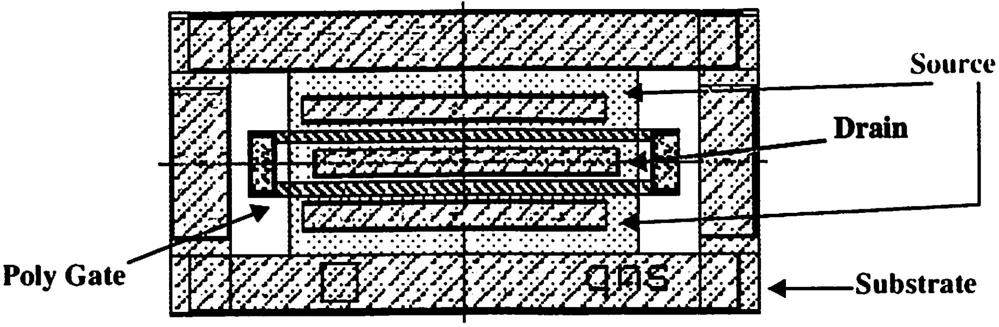


Figure 6-10. Layout of Basic Transistor Cell.

One potential drawback to this method is that the amount of area required to implement the large devices required in this work may be excessive and may cause the the individual cells to see slightly phase-shifted versions of the signals (due to differences in the resistance to gate of the devices in the cells). However, if a low-resistivity metal layer is used to route the input signals, the net difference in the resistance will then be dominated by the via and contact resistance as well as the gate resistance itself, which is the same for each of the individual cells. Therefore the use of this base cell will reduce the impact of the substrate resistance and the R-C time constant of the gate on the performance of the circuit.

6.4.2 Spiral Inductor Design

The on-chip spiral inductors used in this PA were designed using ASITIC[27], a software program written by a UC Berkeley graduate student that provides accurate modelling of on-chip passive components. Each inductor was designed with an optimum Q, in order to maximize the tank impedance seen at the particular nodes of interest. As seen in Figure 6-9, the value of the inductors needed are approximately 1.6-nH and 0.5-nH. In order to maximize the Q, multiple metal layers were used in parallel (to reduce the series resistance of the inductor). There are times when using multiple metal layers is not particularly helpful, but those occur more when designing large inductors and the self-resonant frequency (SRF) of the inductor is close to the operating frequency of the inductor. The SRF is the frequency at which the self-inductance and the parasitic capacitance of the inductor resonate with each other. At frequencies above the SRF, the inductor starts to look capacitive. However, as these inductors are low-value inductors, the SRF is still well above the operating frequency of 1.75-GHz.

The layout of the two load inductors mentioned above is shown in Figure 6-11.

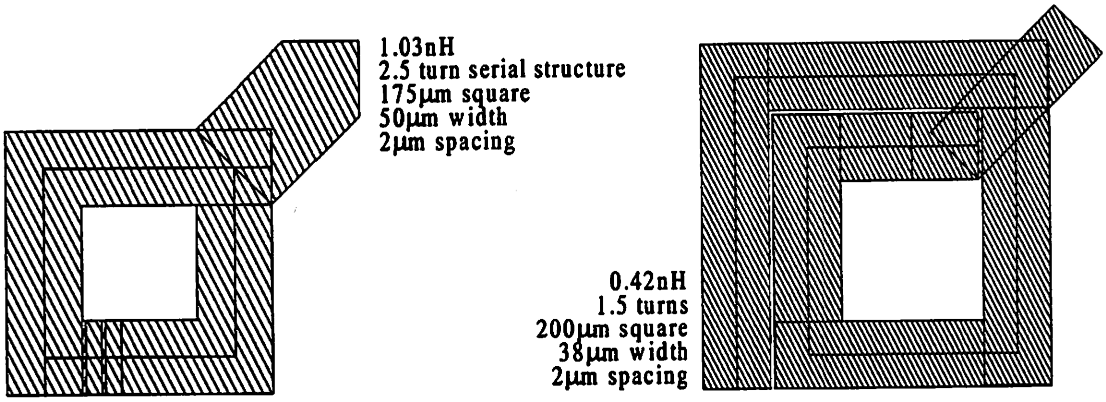


Figure 6-11. Layout Structure of Load Inductors

Two instances of each inductor are needed in the circuit, due to the differential nature of the PA. It is common in the layout of blocks with differential signal paths to use mirror-symmetry to closely match the two signal paths. One problem with mirror-symmetry when dealing with inductors is that the magnetic fields created by two inductors with mirror-symmetry point in the same direction. As a result, the magnetic fields do not cancel, but compound. When placing these inductor on a chip with other circuit blocks which also have many inductors, the magnetic fields generated could interact and cause a degradation in performance. That is precisely the case in this implementation, where the PA will be on the same chip with two frequency synthesizers and a set of RF mixers, both of which use on-chip spiral inductors. As a result, the inductors were not mirrored but rotated 180 degrees. In this way, the magnetic fields generated will cancel each other out so as not to disturb other inductors. Further, any common magnetic field will generate common-mode signals in the inductor, preserving the immunity to common mode signals inherent to differential circuits.

The cascode-node inductors (first described in Section 5.2.5) are also implemented in were each implemented as two inductors in series in order to use the rotated-symmetry layout and increase the immunity of the cascode inductors to common magnetic fields. In the case of the cascode inductors, the Q was less of a concern, as simulations showed that the improvement in efficiency was not all that dependent on the Q of the inductors. As a result, the cascode inductors were designed to first limit the die area and second to optimize the Q. The layout of one of these inductors is shown in Figure 6-12.

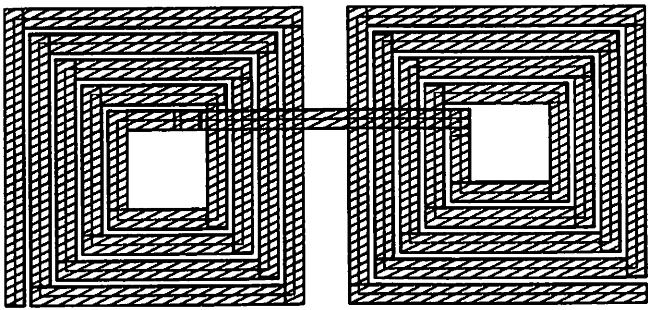


Figure 6-12. Layout Structure of Cascode Inductors

One requirement of using the cascode inductor is that the cascode node must be accessible, which is not the case in standard analog layout practices. In general, the layout of the cascode structure is done as shown in Figure 6-13. The gates of the two

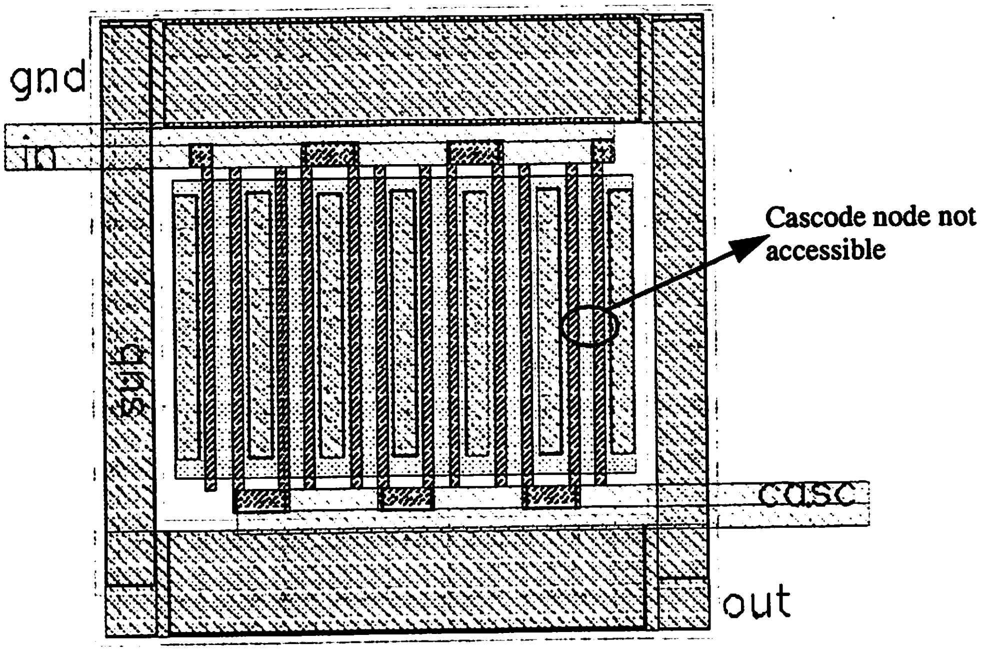


Figure 6-13. Standard Cascode Layout

devices are placed as close together as possible. This method reduces the capacitance on the cascode node, as there is no sidewall capacitance and the junction capacitance at the bottom of the cascode node is also reduced (due to the reduced separation between the two devices). However, the representation shown in Figure 6-13 does not allow an electrical connection to the cascode node, which is needed in order to use the cascode inductor. As a result, the base transistor cell described previously in Section 6.4.1 was used for each of the devices in the cascode structure.

6.4.3 Substrate Coupling

Finally, one issue that needed addressing was substrate coupling. In the integrated implementation, the PA will reside on the same substrate as several other blocks. The PA generates extremely large signals and generates large currents that flow in and out of the substrate. The possibility of these signals corrupting other circuit blocks on the same die is very real, and a few methods for reducing the coupling between blocks (and especially between the PA and the other blocks on the substrate) were used in this work.

First, the entire signal path, from the input of the digital-to-analog converters (DACs) to the output of the PAs, was differential. While the primary reason for using a differential signal path in the PA was the increase in the voltage swing available, there are other positive impacts resulting from the differential implementation. The other blocks retain some immunity to the signals flowing through the substrate, which appear as common mode signals. Further, because the PA is implemented differentially, the PA is now dumping current into the substrate twice every cycle (once for each half-circuit of the differential implementation). As a result, the substrate signals generated by the PA appear to be at twice the frequency of the circuit's operation. With the frequency of the substrate signals being moved away from the operating frequencies of the mixer and especially the frequency synthesizers, the potential for problems such as local-oscillator pulling in the frequency synthesizer is dramatically reduced.

Second, the power supply nodes for each of the blocks on-chip, including the PA, were brought on-chip independently and each supply node was heavily bypassed on-chip to the local electrical ground. Further, the local ground and substrate were tied together so that the ground, substrate, and supply would all move together. The net effect on the circuit itself would thus be minimized, as all the potential movement of the ground, substrate and supply would be common-mode variations, and thus rejected by the differential circuits.

Third, the active devices were flooded with substrate contacts placed as close to the actual devices as possible. The implementation of this concept has already been demonstrated in the base transistor cell used in the PA. The goal of this idea is to present a low-resistance path for the substrate current, so that all the substrate current will be collected by the contacts, rather than flowing through the substrate to other blocks and increasing the overall voltage swing on the substrate.

Finally, the PA itself was completely surrounded by a "wall" of substrate contacts. This substrate wall was used as a secondary "line of defense"; any substrate current not collected by the contacts close to the devices would be collected by this isolating block of contacts, and very little substrate current would get through to the disrupt the other blocks in the circuit. The PA was located in one corner of the chip, and the wall of substrate contacts was approximately 200μm in width.

6.5 Simulation Results

Extensive simulation of the PA was performed in both HSPICE and SpectreRF. Some of the key results include the efficiency, the output power across the transmit band, and the variation of the output power with the change in the tail current of the first stage (the power control mechanism).

Figure 6-14 shows some of they key terminal voltages along the signal path.

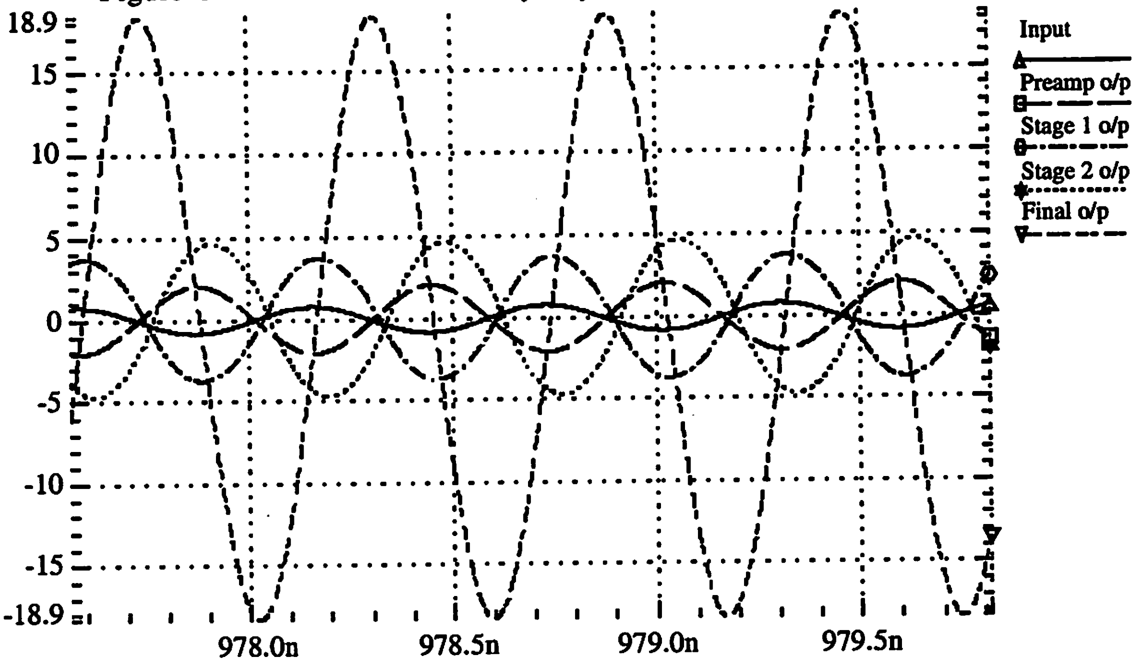


Figure 6-14. Interstage and Final Output Voltages

The preamplification stage described in Section 6.3.7 is driven with a sinusoid with a 0.4-V amplitude (zero-peak). The voltages shown are the differential voltages at the input of the first, second, and final stages. The output power of the PA is shown versus frequency in Figure 6-15. The power is plotted in dBm, which is the most common units in which RF PA power is described. As can be seen, the output power is extremely constant over the frequency of interest. Also shown in the figure is the PA efficiency over the same range. This efficiency number is the overall efficiency. In this case, the overall efficiency is essentially equivalent to the PAE; the power required to drive the PA is minimal, as the input presented to the previous stage is entirely capacitive. The efficiency of the PA is approximately 40% over the transmit band. While this is slightly below the goal of 50% that was initially set, this is primarily due to the fact that the PA was designed for an output power of 1.7W instead of a level closer to 1W. In fact, before the desired peak output power level of the PA was increased to 1.7W, a PA delivering just over 1-W was designed with an efficiency of 48%, very close to the desired 50% goal. However, once the peak output power goal was raised to 1.7W, the optimum output resistance was reduced and the PA needed to generate even more current, reducing the efficiency to about 40%.

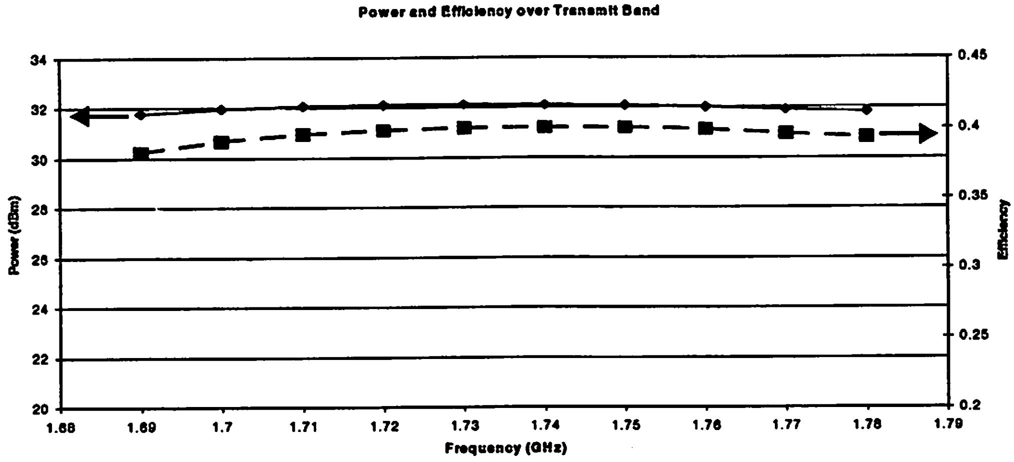


Figure 6-15. Output Power and Efficiency vs. Frequency in the Transmit Band

Figure 6-16 shows the PA output power as a function of the tail current in the first stage of the PA. As expected, the PA output power is related to the tail current; as

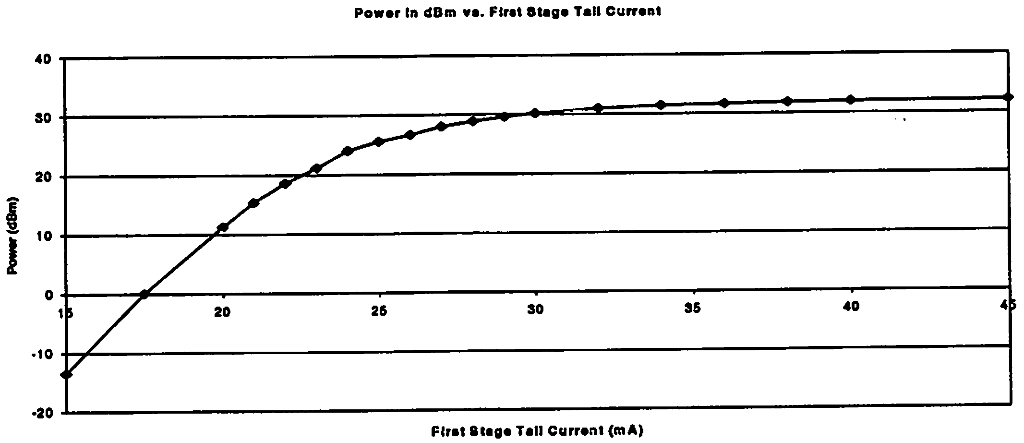


Figure 6-16. PA Output Power vs. First Stage Tail Current.

the tail current rises, so too does the output power. It should be noted, however, that even though the relationship looks linear for the middle portion of the graph, it is not. The power is again plotted in dBm, so the graph is actually a log-linear plot, and therefore the relationship between input and output is non-linear.

6.6 Stand-alone Test Chip

As stated in the opening to this chapter, the PA was also implemented as a stand-alone test chip. The test-chip included the output stage as well as the first and second stage described previously. The preamplifier discussed in Section 6.3.7 was not included in the test-chip.

The layout of the PA was copied exactly from the overall transceiver block, and the extra available pads were used primarily for connections to electrical ground. In the implementation of the test chip, it was possible to make use of the extra pads available to enhance the performance of the PA. However, as one of the driving factors behind the implementation of the test-chip was the ability to better understand the functionality of the PA that was in the signal path of the integrated transceiver, no layout modifications were made to the class-C PA.

6.7 Conclusion

This chapter has described the circuit implementation of a CMOS Class-C PA, using the techniques described in earlier chapters. The simulated PA was able to deliver over 1.6-W of power to a 50 ohm load, with a PAE of 40%. The next chapter will discuss the experimental set-up, including the evaluation board designed for testing the PA, as well as the experimental results.

7 Experimental Results

In Chapter 6, the actual circuit implementation of the Class-C PA was described. In this section, experimental results from both the stand-alone test chip as well as the integrated version of the PA will be presented. This chapter will begin with a discussion of the test board for the stand-alone test-chip as well as a mention of aspects of the test board for the fully integrated transceiver that relate to the Class-C PA. The actual experimental results will follow. A die photo of the stand-alone die is presented in Figure 7-1. The individual stages and the signal path are displayed on the figure.

7.1 Test Board

The design of the test board for the stand-alone PA was quite straightforward. A photo of the die attached to the board with its bond wires is shown in Figure 7-2. As stated earlier in Section 5.2.3, the die was attached directly to the board in the Chip-On-Board (COB) fashion. The die-attach area is highlighted in the figure. As shown, the input to the PA is brought on board from the left and converted to differential signal using a MuRata balun (the balun is not shown in the picture; the inputs to the chip are

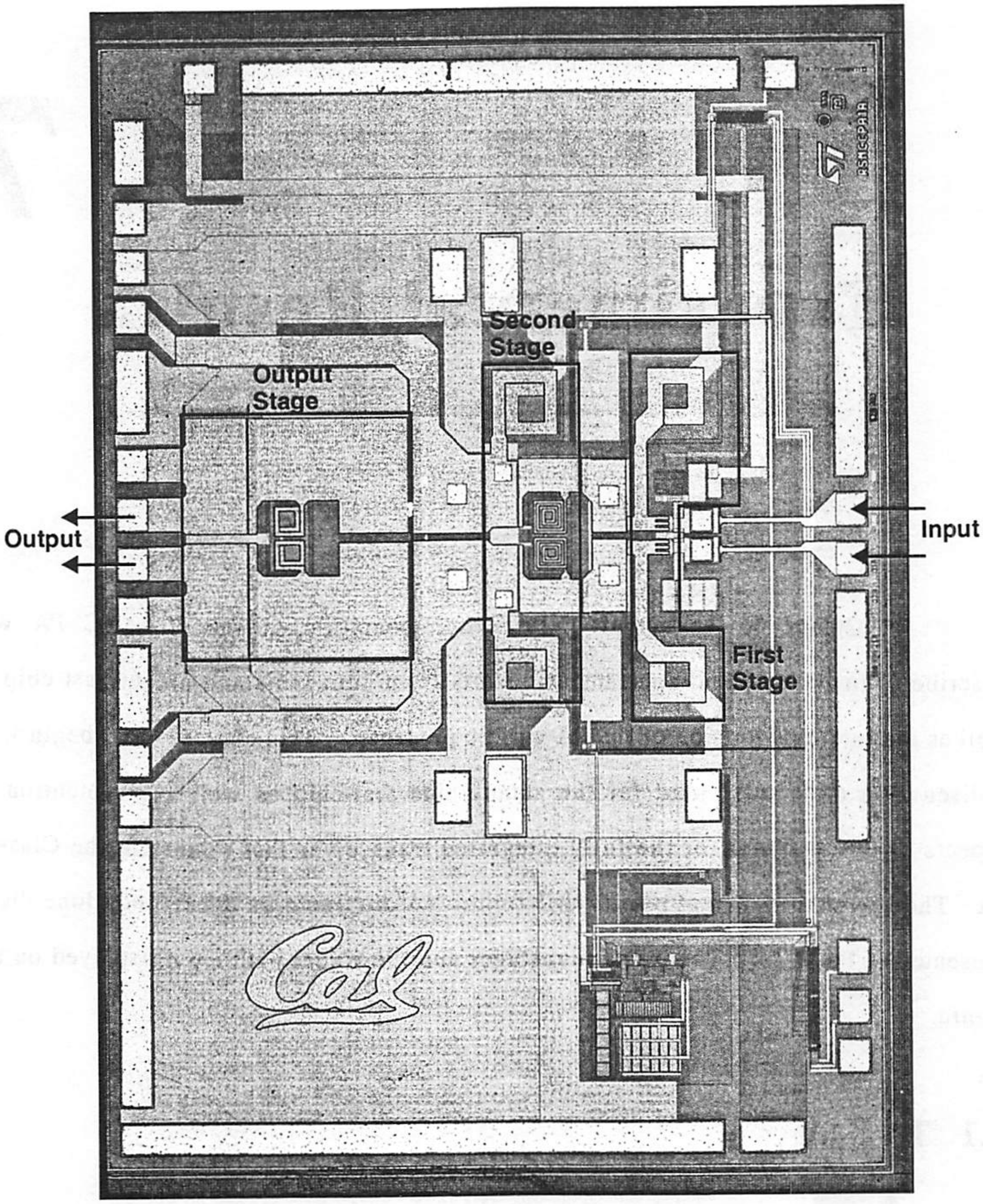


Figure 7-1. Stand-Alone PA die photo

already differential). The differential signal is then applied to the PA. The output of the PA comes out from the right of the board. In this board implementation, the PA outputs are available individually (no power combination has been performed). This was done in order to remove the uncertainty in the performance of any balun or power combiner. The other required elements on the board include a regulated current source,

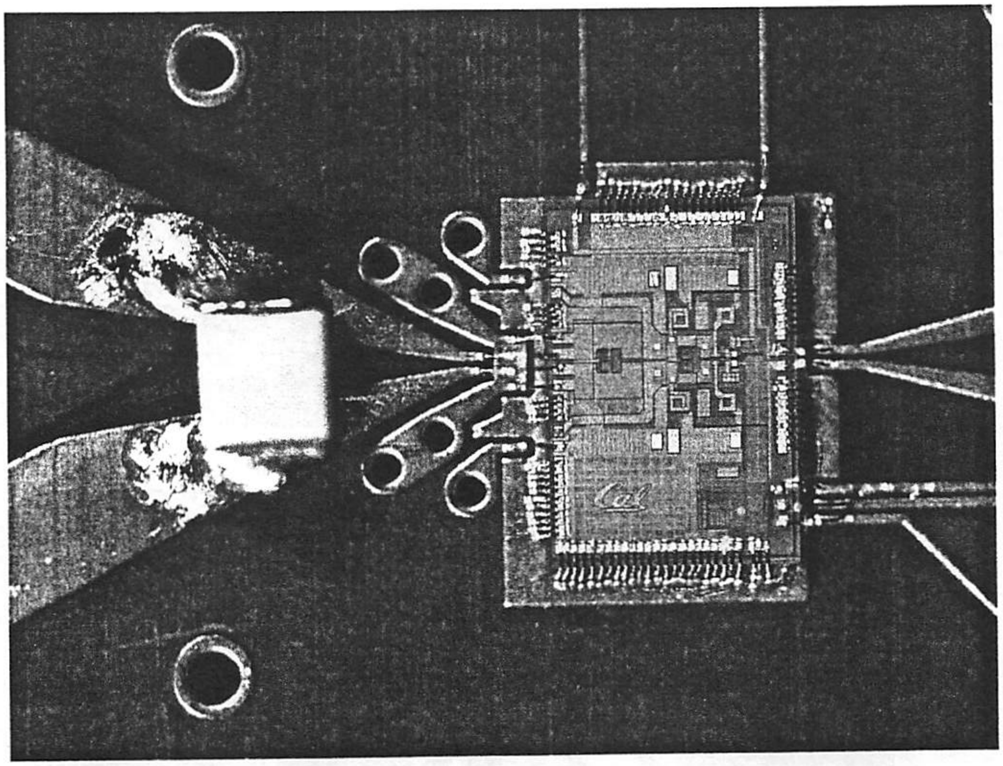


Figure 7-2. Stand-alone PA on test board (with bond wires)

bypass capacitors, and headers required to drive the shift register signal on chip. Two separate power supply voltages were used: one which powered the first two stages of the PA (and, in the integrated version, the preamplifier stage as well), and one which powered the final output stage. This was done for two reasons: first, the supplies on-chip were separated in this manner, and second, the current consumption of the final stage could be viewed independently of the earlier stages.

A second implementation of the board did contain a microstrip balun, which was based on a design first used previously[16]. This microstrip balun was designed to minimize the loss in the combining process, and included the bondwire inductors in the design process to accurately present the proper load to the output of the transistors on-chip. The simulated loss of the balun was approximately 0.6-dBm. The board for the integrated version of the transmitter, including the PA, also used the microstrip balun. The entire transmitter attached tot he board is shown in Figure 7-3(a); the microstrip balun and the edge of the die used by the class-C PA is shown in closer detail in Figure 7-3(b). The portion of this board corresponding to the output of the class-C PA matches the evaluation board of the test-chip. This was done again so that the performance of the test-chip would closely match that of the PA in the integrated version.

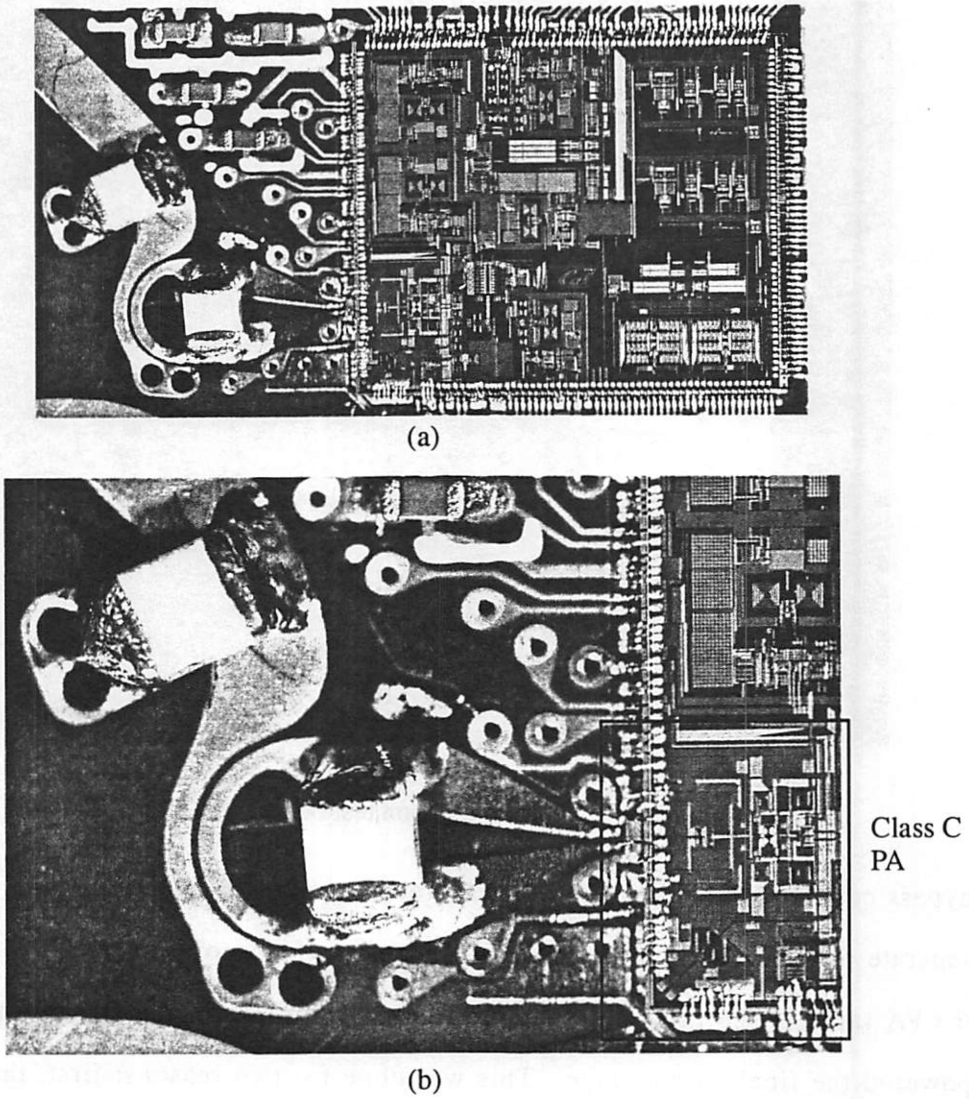


Figure 7-3. Integrated Transceiver Test Board

7.2 Experimental Test Setup

The test-setup for evaluating the PA was slightly different for the stand-alone test-chip than for the integrated version. The test-setup for the stand-alone test-chip is shown in Figure 7-4. The signals applied to the board were just the two power supplies

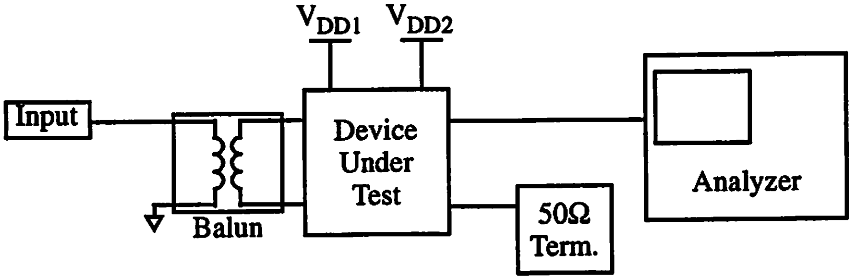


Figure 7-4. Test-setup for Stand-alone test-chip

and the input drive signal (supplied from a Rohde&Schwarz signal generator). The output was fed into a spectrum analyzer. In Figure 7-4, the differential outputs from the chip are available independent from each other, and as shown, each side is individually tested. In certain tests, a power combiner was also used to combine the two output signals and examine the full output together.

In the case of the integrated transceiver, the input to the PA was not easily accessible. Therefore, in order to test the PA, the entire transmit chain needed to be functional. The input signal was in the form of two 10-bit digital baseband signals which were fed into the DAC, which was generated by an HP pattern generator. That signal went through the transmit chain, first being converted to an analog signal by the DAC, then filtered and then frequency translated to the RF carrier frequency using a two-step mixing process which included the Harmonic-Rejection Mixers described in Section 2.2. The on-chip frequency synthesizers synthesize the LO signals used by the mixers to perform the frequency translation. The output of the mixers drives the preamplification stage described in Section 6.3.7. The output of the PA was fed into a spectrum analyzer.

7.3 Results

This section will be divided into two sections. The first will cover the experimental results from the stand-alone test-chip, and the second will cover the results from the PA included with the integrated transceiver. In the second case, some of the results will be compared to the output of the mixers (before the PA), in order to examine any potential degradation in the performance of the transmitter.

7.3.1 Stand-alone Test-chip

The stand-alone PA test-chip was initially tested in order to obtain an accurate measure of the performance of the PA without any factors external to the operation of the PA disrupting the measurements. It was apparent from the initial results that the gain through the PA was less than expected from the simulations performed prior to the chip's fabrication. Due to the fact that several methods of increasing the gain were available, the overall gain of the PA was raised and eventually the desired output power goal (> 1W power output) was met. However, these methods of increasing the gain of the PA had the undesired side-effect of reducing the efficiency, as more current was consumed in order to increase the output power.

The primary reason for the reduction in the gain was believed to be the quality factor (Q) of the on-chip spiral inductors. On-chip inductors are used at the output of the first and second stages in order to tune out the load capacitances and peak the gain at the frequencies of interest. As discussed previously in Chapter 5, the Q of the inductor determines the impedance level of the corresponding L-C tank. As the Q decreases, so too does the impedance level, resulting in a lower voltage gain for a given output current. Unfortunately, the individual inductors were not available for characterization, so exact Q numbers were not obtained. However, after the chip was fabricated, a more accurate version of the ASITIC simulator became available. The simulated Q of the inductors in the older version of ASITIC as well as the newer, more accurate version, are shown in Table 1. The "Final Q" column lists the inductor Q values resulting from the newer version of ASITIC. As can be seen, the change in Q is

Table 1: Simulated Inductor Values

Inductor	Value	Original Q	Final Q
First Stage Output	1.03n	6.5	4.5
Second Stage Output	0.42n	4.9	2.3
Second Stage Cascode	2.6n	4.1	2.0
Third Stage Cascode	0.54n	3.3	1.7

quite dramatic, and such a change will have a large impact on the performance of the overall circuit.

In order to counter the impact of the reduced Q of the spiral inductors, two methods were used for increasing the gain. First, the tail current in the first stage was increased, and the bias voltage at the gate of the second stage was increased as well. Both of these methods serve to increase the gain of the overall PA. The increase in current in the first stage increases the current to the reduced load impedance. The increase in the bias voltage at the gate of the second stage also increases the current generated by the second stage, further increasing the gain. The third stage was left unchanged; any increase in the current level of the third stage would dramatically increase the overall current consumption of the PA, dramatically reducing the efficiency. While the increasing current consumption in the previous stages will also reduce the efficiency, because of the current levels in those stages are low compared with the final stage, the reduction in efficiency will not be as dramatic as it would be if the current level in the final stage were increased.

The results from the stand-alone PA are presented. The stand-alone PA is driven with a signal whose magnitude is 10dBm. The actual power driven into the chip was about 9dBm due to a 1-dB insertion loss from the MuRata balun. With the increased current levels in the first and second stages, the output power over the frequency band of peak output power is shown in Figure 7-5. It should be noted that the

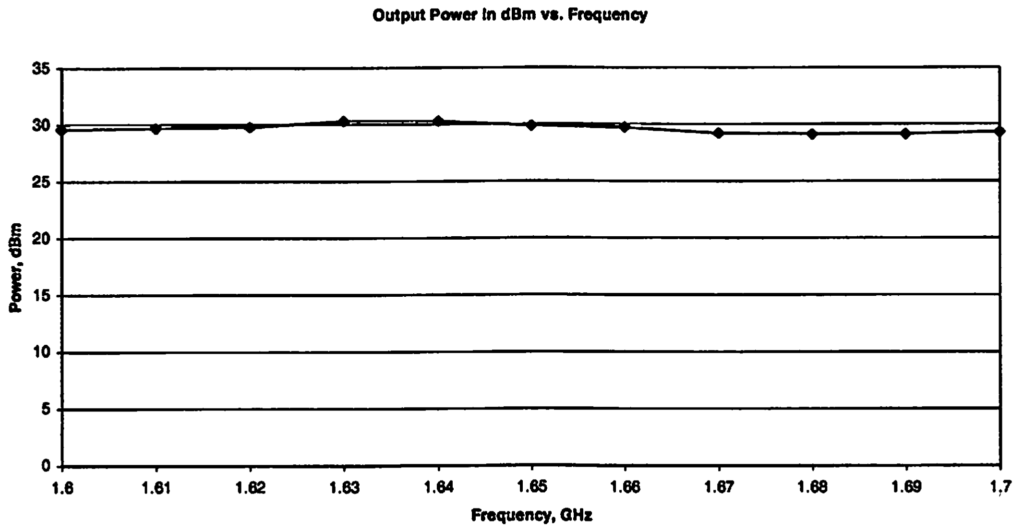


Figure 7-5. Measured Output Power vs. Frequency for Stand-alone PA

peak output power occurred at about 100 MHz away from the predicted peak output power. This is likely due to a slight deviation in the load inductor values than predicted, or a slightly larger capacitance in the signal path than predicted. The error is on the order of 6%, which can easily be corrected for. The measured efficiency over this range of frequencies is shown in Figure 7-6. The drain efficiency and PAE of the

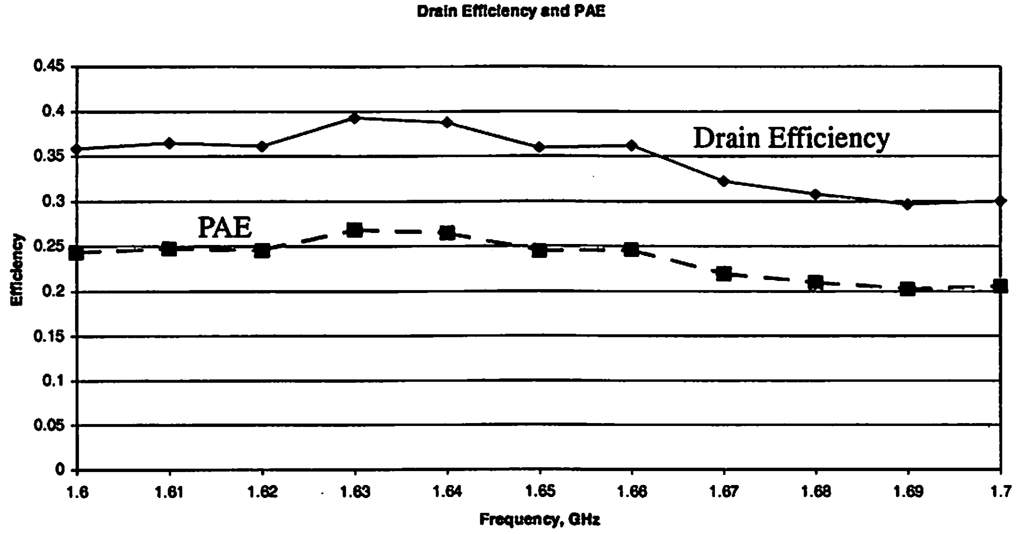


Figure 7-6. Efficiency vs. Frequency for Stand-alone PA

PA are shown in the figure. The output power as a function of the tail current in the first stage is shown in Figure 7-7.

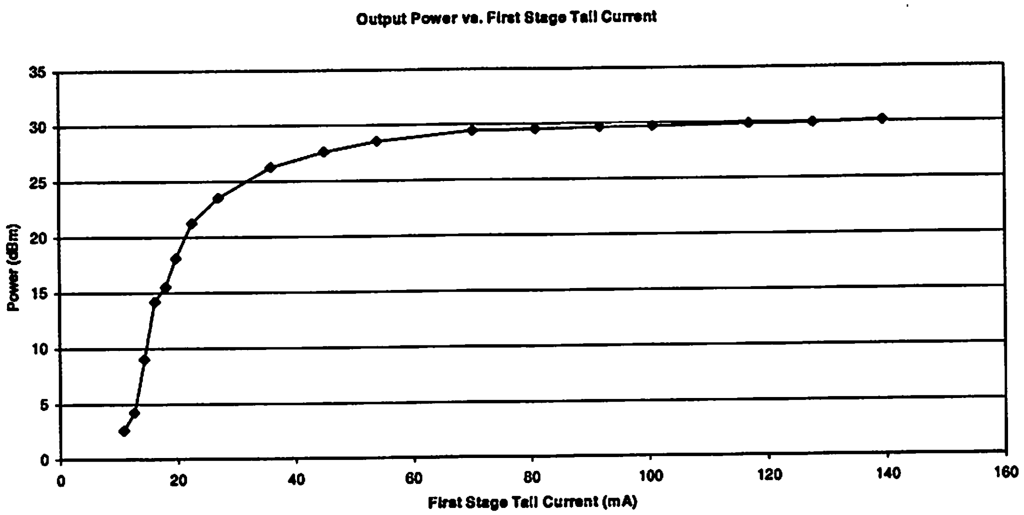


Figure 7-7. Output Power vs. First Stage Tail Current for Stand-alone PA

In order to verify that the dramatically lower inductor Q was a plausible cause of the reduced gain through the PA, HSPICE simulations were performed using the new values of the inductor Qs presented in Table 1. The output power and efficiency resulting from these new simulations is shown in Figure 7-8. As can be seen, the power

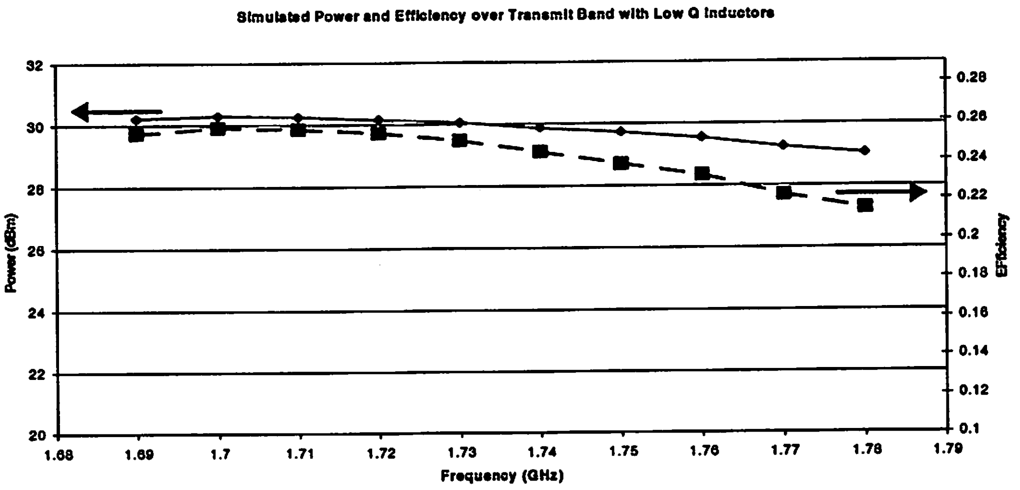


Figure 7-8. Simulated Output Power and Efficiency with Reduced Q Inductors

and efficiency are much more in line with the actual power and efficiency that was obtained from the stand-alone PA. As a result, the hypothesis that the PA performance was degraded by the low Q of the spiral inductors is a reasonable one.

The output of the PA can also be viewed in the frequency domain. The output of the PA with a single-tone input is shown in Figure 7-9. The peak is at the input

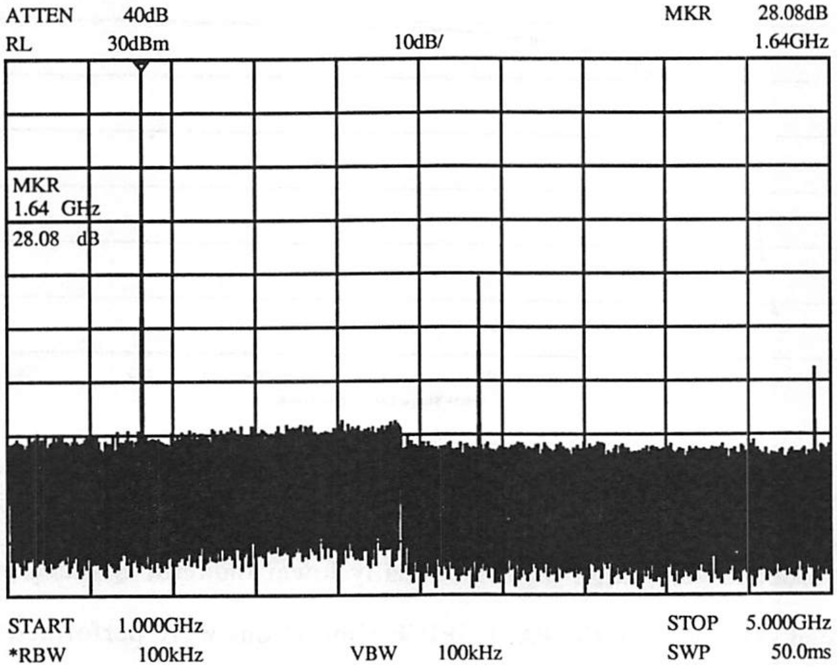


Figure 7-9. Frequency Domain output of Stand-Alone PA driven by Single-Tone Input

frequency of 1.64 GHz. Due to the use of a power combiner, there was about 2dB of loss between the chip output and the spectrum analyzer. Furthermore, one important measure of the performance of a PA is the output of the PA under the input of a GMSK modulated signal. In the DCS1800 specification, a spectral mask is given which restricts where the energy of the signal can be transmitted. Figure 7-10 shows the output of the PA when driven by a GMSK modulated signal. As can be seen the output of the PA is well within the output spectral mask supplied by the DCS1800 specification.

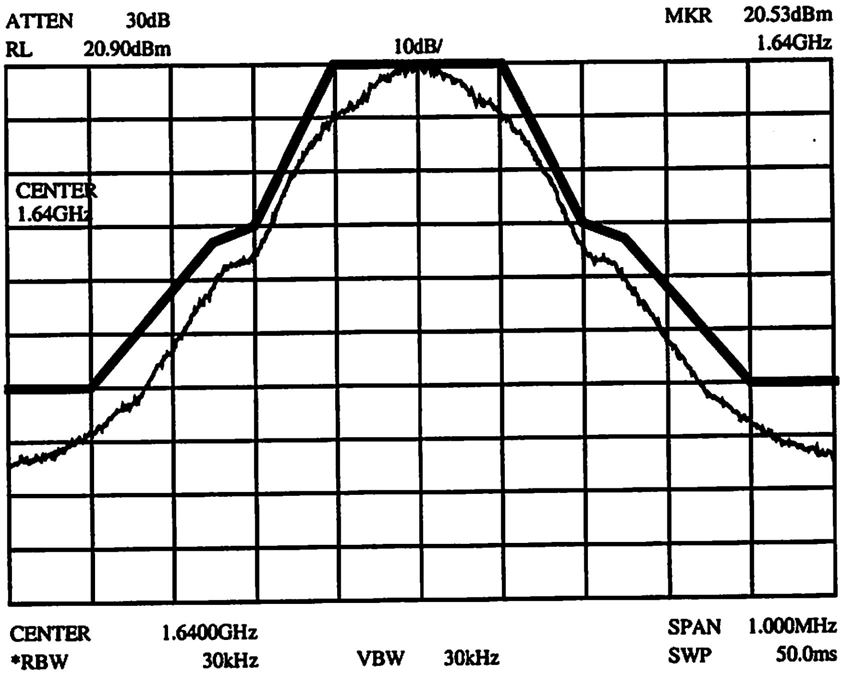


Figure 7-10. Output Spectral Mask of Stand-Alone PA

7.3.2 Integrated PA

A die photo of the integrated transmitter, including the Class C PA, is shown in Figure 7-11. The low-Q problem that plagued the stand-alone PA had a larger impact on the operation of the PA included in the integrated transmitter. In this implementation, not only was the gain reduced in the PA itself, but the gain in the previous stages of the transmitter was also reduced, as previous stages also used spiral inductors. This dramatically reduced the magnitude of the input signal driving the PA, reducing the output level further. Due to this reduced signal level, it was necessary to further increase the gain of the PA in order to compensate. In this case, not only were the tail current in the first stage and the bias of the gate of the second stage maximized, so too was the bias of the gate of the third stage. Even still, the peak output of the PA still did not meet the 30dBm output power level required by DCS1800; the maximum power output was approximately 25dBm. However, some results are presented here to show the impact of the PA on the output signal. Shown in Figure 7-12 is the output of the transmitter when driven with a single tone. The output of both the mixers (driven off-

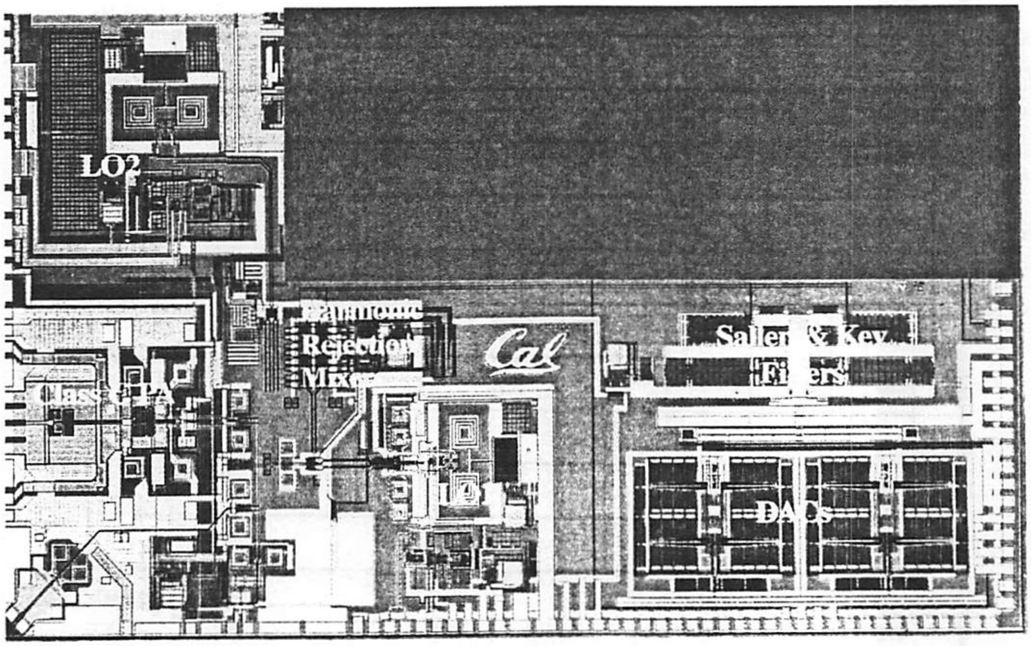
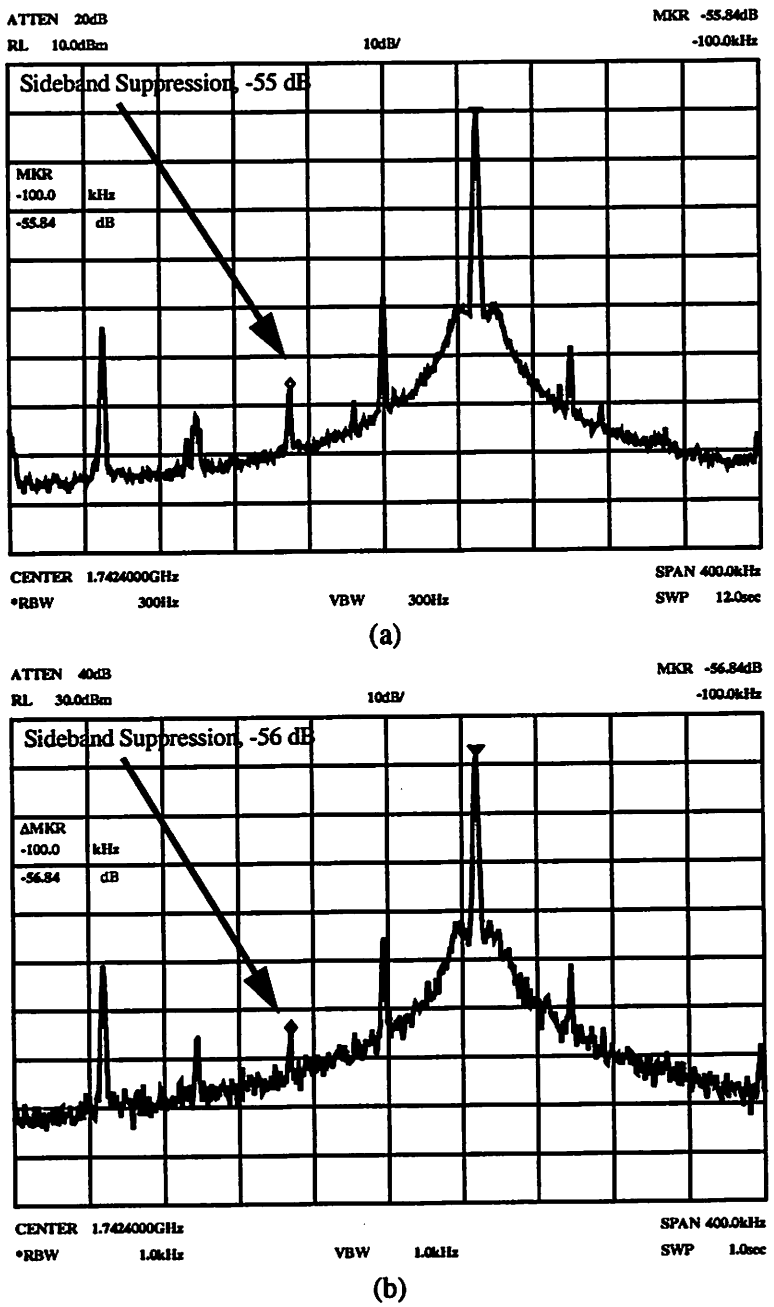


Figure 7-11. Integrated Transceiver Die Photo

chip with a test buffer) and the PA are both shown. The PA has very little impact on the close in performance of the transmitter; the LO feedthrough and the lower sideband are virtually unchanged, and the IM3 is only slightly larger. A second test that was performed investigated the noise performance of the PA at a 20-MHz offset from the center frequency. In DCS1800, this is quite an important specification, as 20 MHz is the closest distance between the transmit and receive bands, and there are stringent limits on the amount of transmission allowed by the transceiver in the receive band. The noise performance at 20MHz away from the carrier was seen to be about -126dBc/Hz at the output of the mixers and -121dBc/Hz at the output of the PA. The wideband noise performance at the output of the PA is not dramatically affected by the use of the PA. It is expected that this increase is due to the increased impact of substrate coupling on the performance of the transmitter (especially the frequency synthesizers) when the PA is on. Further, it is quite possible that the noise performance might be better if the PA was working as originally intended. As tested, the current flowing through the integrated PA was at its maximum; that is, the biases of both the second and third stages were at their highest point, maximizing the current flowing through those stages, and



(a)

(b)

Figure 7-12. Single Tone Output (a) Before and (b) After Integrated PA

increasing the current flowing in the substrate of the device. If the substrate current were to be reduced, it is possible that the 5-dB increase in the wideband noise would be reduced.

Due to the dramatically lower power levels than were expected, the integrated PA was not as fully tested as the stand-alone PA. The integrated PA was primarily tested to investigate its impact on the output signal when compared with the signal prior to the PA. The results discussed above were the primary results obtained from the PA on the integrated test-chip.

7.4 Conclusions

In this chapter, experimental results from the stand-alone and integrated versions of the prototype class-C PA discussed in Chapter 6 were presented. Due to the lower than expected on-chip spiral inductor Qs, the performance of the PA was not quite what was predicted by the simulations performed prior to the fabrication of the prototype. However, an output power level of greater than 30dBm was still achieved by the stand-alone PA, and the stand-alone PA met the spectral mask required by the DCS1800 specification under the peak power condition. A list of the key performance metrics of the PA is presented in Table 2. Furthermore, the PA also achieved the desired characteristic of having the output power level be able to be controlled by the magnitude tail current in the first stage. The integrated PA did not meet the required

Table 2: Power Amplifier Performance

V_{DD}	3.3V
Stand-alone PA:	
Power In	10dBm
Peak Power Out	30.33dBm
Peak Drain Efficiency	39.3%
Peak PAE	26.9%
Meets GSM Spectral Mask	Yes
Integrated PA:	
Peak Power Out	25dBm
Single Sideband Suppression Change from Pre-PA Measurement	-56.84dBc -1dB
Wideband (20 MHz) Noise Change from Pre-PA Measurement	-121dBc/Hz 5dB

output power level, unfortunately, but the impact of a non-linear PA on the transmitted signal was minimal.

8

Conclusions

If the goal of a single-chip CMOS radio transceiver for cellular wireless communications systems is to be achieved, it is clear that two critical goals must be accomplished in the implementation of these transceivers: first, a high-power (> 1W) power amplifier (PA) in must be implemented in CMOS, and second, that PA must be included on the same substrate as the rest of the transceiver circuitry. Furthermore, the PA must have a reasonable efficiency that it can be somewhat competitive with PAs in other technologies.

There are several obstacles which make the implementation of such a PA extremely difficult. The more common classes and architectures used in PA design, such as Classes A and AB, are inherently lower-efficiency classes. Furthermore, the use of sub-micron CMOS processes increases the difficulty, as technology limitations specific to CMOS processes, such as limited voltage swing due to oxide breakdown and poor transconductance, complicate the design process.

In this work, a 1.75-GHz CMOS PA was designed and implemented. In order to maximize the efficiency, a Class-C architecture was used. Key research contributions from this work are summarized here:

- A simplified design methodology was developed for the design of Class-C PAs in CMOS technologies. The design methodology predicts the drain current of the device using a fourier series analysis of the drain current as well as iteration in order to approximate when the device is in saturation and when it is in triode.

- Several different circuit techniques were used in order to overcome the limitations inherent to CMOS technologies, such as limited voltage swing and poor transconductance. New techniques, including the use of cascode inductors and the use of a modified tuning method in order to peak the impedance and therefore the gain at the input to the final drive stages, allowed the use of extremely large devices which facilitated the high output power required for the prototype designed in this work.

- A 1.7-W, 1.75-GHz PA has been designed in this work in a 0.35–μm CMOS process, using a Class-C architecture. Experimental results indicated that the PA actually delivered 1-W of output power to the load, due to lower-than-expected Qs from the on-chip spiral inductors.

- A 40% efficient, 1.7-W, 1.75-GHz, Class-C PA was designed in a ST Microelectronics 0.35-μm, 5-metal, double poly process. Unfortunately, due to an over-estimation of the quality factor (Q) of the on-chip spiral inductors used in the design, the stand-alone PA only provided just over 1W of output power with a peak efficiency of 27%.

- The Class-C PA was also included in an integrated, single-chip CMOS transmitter, which integrated all the blocks in the transmitter chain from the digital-to-analog converters through the PA along with two frequency synthesizers. The power output of the PA in the transmitter was even less than that of the stand-alone version, due to the reduced Q of the on-chip spiral inductors, but the PA did not appreciable distort the transmitted signal.

- This PA is the first example of a Class-C PA implemented in a standard CMOS process which delivers 1W of power at RF frequencies.

The goal of this work was primarily to verify the feasibility of implementing PAs for high-power mobile applications in CMOS technologies using architectures that support high-efficiency, and to that end, it succeeded. While the efficiency of the implementation was low, the simulated efficiencies were significantly higher, and the post-implementation simulations using the low-Q inductors mirrored the performance of the prototype.

If this work were to be continued, the use of newer sub-micron CMOS processes which hold the promise of passive components (such as on-chip spiral inductors) with Qs of 5-10 would greatly benefit PA design, and a 1.7-W PA with an efficiency closer to 50% could be implemented using the techniques described in this work. However, as lithography and sub-micron processes become better, one overwhelming problem continues to plague the PA designer: the oxide-breakdown voltages are being reduced along with the minimum device sizes. Cutting-edge processes (such as 0.18mm and 0.1mm) usually have "output-driver" devices which have thicker oxides and can withstand larger voltages, but these devices are usually equivalent to 0.35mm devices. Therefore, while these newer processes will hopefully give better passive components, the devices themselves available to the PA designer will not be significantly different from the devices used in this design unless realistic methods of using low-voltage devices while delivering large amounts of power can be identified.

9 References

[1] P.R. Gray and R.G. Meyer, "Future Directions of Silicon IC's for RF Personal Communications," *Conference Proceedings, 1995 Custom Integrated Circuits Conference*, May 1995, pp. 93-90.

[2] J.A. Weldon, J.C. Rudell, L. Lin, R. S. Narayanaswami, et. al., "A 1.75-GHz Highly-Integrated Narrow-Band CMOS Transmitter with Harmonic-Rejection Mixers," *2001 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, February 2001, pp. 160-161,442.

[3] J.A. Weldon, Ph.D. Thesis, University of California, Berkeley, to be published.

[4] D. Yee, et. al., "A 2-GHz Low-Power Single-Cip CMOS Receiver for WCDMA Applications," ESCARP 2000.

[5] D. Shaeffer, et. al., "A 115-mW, 0.5-mm CMOS GPS Receiver with Wide Dyname-Range Active Filters," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 12, December 1998, pp. 2219-31.

[6] J.C. Rudell, et. al., "A 1.9GHz Wide-Band IF Double Conversion CMOS Receiver for Cordless Telephone Applications," *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 12, December 1997, pp. 2071-88.

[7] L. Lin, L. Tee, P.R. Gray, "A 1.4GHz Differential Low-Noise CMOS Frequency Synthesizer using a Wideband PLL Architecture," *2000 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, February 2000, pp. 204-5,458.

[8] R.S. Narayanaswami, *The Design of a 1.9GHz, 250mW CMOS Power Amplifier for DECT*, Master's Report, University of California Berkeley, May 1998.

[9] T.H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, Cambridge, United Kingdom, 1998.

[10] H.L. Krauss et al, *Solid State Radio Engineering*, Wiley, New York, 1980.

[11] J. S. Kenney and A. Leke, "Power Amplifier Spectra Regrowth for Digital Cellular and PCS Applications," *Microwave Journal*, October 1995, pp. 74-92.

[12] P.R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits, Third Edition*, John Wiley & Sons, New York, 1993.

[13] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*, Artech House, Boston, 1999.

[14] N. Sokal and A. Sokal, "Class E-A New Class of High-Efficiency Tuned Single-Ended Switching Power Amplifiers," *IEEE Journal of Solid State Circuits*, Vol. SC10, June 1975, pp. 168-76.

[15] D. Su, W. McFarland, "A 2.5-V, 1-W Monolithic CMOS RF Power Amplifier," *Proceedings of the IEEE 1997 Custom Integrated CIrcuits Conference*, May 1997, pp 189-92.

[16] K.C. Tsai and P. R. Gray, "A 1.9GHz CMOS Class E Power Amplifier for Wireless Communications," *IEEE Journal of Solid-State Circuits*, Vol. 34, No. 7, July 1999, pp. 962-70.

[17] T. Sowlati, et. al., "Low Voltage, High Efficiency GaAs Class E Power Amplifiers for Wireless Transmitters,", *IEEE Journal of Solid State Circuits*, Vol. 30, October 1995, pp. 1074-80.

[18] A. Bateman, D. Haines, and R. Wilkinson, "Direct Conversion Linear Transceiver Design," *Fifth International Conference on Mobile Radio and Personal Communications*, London, UK, 1989, pp. 53-6.

[19] A. Lohtia, P. A. Goud and C. G. EngleField, "Adaptive Digital Linearization of RF Power Amplifiers," *Canadian Journal of Electrical & Computer Engineering*, Vol. 20, April 1995, pp. 65-71.

[20] A. Mansell and A. Bateman, "Practical Implementation Issues for Adaptive Predistortion Transmitter Linearization," *IEE Colloquium on Linear RF Amplifiers and Transmitters*, London, UK, 1994, pp. 5/1-7.

[21] P.B. Kenington, R.J. Wilkinson, and J.D. Marvil, "Broadband Linear Amplifier Design for a PCN Base-Station," *Proceedings of the 41st Vehicular Technology Conference*, St. Louis, May 1991, pp. 155-60.

[22] S. Wolf, *Silicon Processing for the VLSI Era Volume 3: The Submicron MOSFET*, Lattice Press, Sunset Beach, California, 1995.

[23] T.H. Distefano, M. Shatzkes, "Impact Ionization model for dielectric instability and breakdown," *Applied Physics Letters*, Vol. 25 no. 12, Dec. 1974, p. 685.

[24] I.C. Chen, S. Holland, C. Hu, "Electrical Breakdown of Thin Gate and Tunneling Oxides," *IEEE Transactions on Electron Devices*, p. 413, Feb. 1985.

[25] D.R. Wolters, A.T.A. Zeegers-van Duijnhoven, "Breakdown of Thin Dielectrics," *Ext. Abs. Mtg. of Electrochem. Soc.*, Spring 1990, p. 272.

[26] E. R. Minami, S.B. Kuusinen, E. Rosenbaum, P.K. Ko, C. Hu, "Circuit-Level Simulation of TDDB Failure in Digital CMOS Circuits," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 8, No. 3, August 1995.

[27] A. Niknejad, *Analysis, Design, and Optimization of Spiral Inductors and Transformers for Si RF ICs*, Master's Report, University of California, Berkeley.

[28] S. Chu, K.W. Chew, W.B. Loh, Y.M. Wang, B.G. Onn, Y. Ju, J. Zhang, K. Shao, "High Quality Factor silicon-integrated spiral inductors achieved by using thick top metal with different passivation schemes," *Proceedings of the 2001 International Symposium on VLSI Technology, Systems, and Applications*, April 2001, pp 154-7.

[29] J.N. Burghartz, D.C. Edelstein, K.A. Jenkiin, Y.H. Kwark, "Spiral inductors and transmission lines in silicon technology using copper-damascene interconnects and low-loss substrates," *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, No. 10, Oct. 1997, pp. 1961-8.

[30] A. Niknejad, *Analysis, Simulation, and Applications of Passive Devices on Conductive Substrates*, Ph.D. Thesis, University of California, Berkeley, 2000.

[31] J.H. Huang, Z.H. Liu, M.C. Jeng, P.K. Ko, C. Hu, "A Robust Physical and Predictive Model for Deep-Submicrometer MOS Circuit Simulation," *Proceedings of the 1993 Custom Integrated Circuits Conference*, May 1993, pp. 14.2.1-14.2.4.

[32] S.C. Jain, P. Balk, "A unified analytical model for drain-induced barrier lowering and drain-induced high electric field in a short-channel MOSFET," *Solid-State Electronics*, vol. 30, no. 5, May 1987, pp. 503-11.

[33] Y.-G. Lee, H.-Y. Lee, "Novel high-Q bondwire inductor for MMIC," *Technical Digest, International Electron Devices Meeting*, 1998, pp 548-51.

[34] J.-J. Ou, X. Jin, I. Ma, C. Hu, P.R. Gray, "CMOS RF Modeling for GHz Communication IC's," *Digest of Technical Papers, 1998 Symposium on VLSI Technology*, June 1998.