Efficiency Enhancement Techniques for CMOS RF Power Amplifiers

Naratip Wongkomet
Paul R. Gray

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-74
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-74.html

May 19, 2006

Efficiency Enhancement Techniques for CMOS RF Power Amplifiers

by

Naratip Wongkomet

B.Eng. (Chulalongkorn University, Thailand) 1998
M.S. (Georgia Institute of Technology) 1999

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
Professor Ali Niknejad
Professor Paul K. Wright

Spring 2006

The dissertation of Naratip Wongkomet is approved:

_____

Professor Paul R. Gray, Chair Date

_____

Professor Ali Niknejad Date

_____

Professor Paul K. Wright Date

University of California, Berkeley

Spring 2006

Efficiency Enhancement Techniques for CMOS RF Power Amplifiers

by

Naratip Wongkomet

Abstract

Efficiency Enhancement Techniques for CMOS RF Power Amplifiers
by
Naratip Wongkomet

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Paul R. Gray, Chair

Growth in the wireless communication market in recent years has been driving the demand for higher integration of CMOS wireless transceivers in order to achieve lower cost, smaller form factor, and more functionalities. Much recent research effort has demonstrated the feasibility to integrate most transceiver building blocks into a single CMOS die. One of the few remaining blocks that has yet to be successfully integrated is the Power Amplifier (PA). The PA is usually the last active building block in a radio transmitter. Its function is to amplify the signal power up to the required level before it can be transmitted into the air. Due to several limitations of CMOS technology, designing a linear and efficient PA is a challenging task.

A property shared among most PAs is that the maximum power efficiency is achieved only when the PA is transmitting peak output power. Efficiency degrades dramatically as output power decreases. Under typical operating conditions, the PA transmits well below its peak output power, therefore the

effective efficiency is much lower than the maximum value. This thesis applies a concept first developed in the vacuum tube era by William H. Doherty to improve amplifier efficiency over a wide range of output power to the CMOS PA problem. Several circuit techniques are also explored in order to optimize efficiency and linearity of CMOS PA, and to allow a high level of integration.

A highly integrated PA prototype was designed in a 0.13μm CMOS technology. It is designed to operate in the cellular DCS1800 band, which has the transmit frequency between 1710MHz and 1785MHz. With GMSK modulated signal, the prototype achieves +31.8dBm output power with 36% power-added efficiency (PAE). The PAE stays above 18% over 10dB range of output power. The PA also meets the GSM/EDGE spectral mask requirement at +25dBm output power with 13% PAE.

_____

Paul R. Gray, Chairman of the Committee

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Advancements in wireless devices in the past two decades have enabled explosive growth in the wireless communication market. With the advent of radio frequency (RF) CMOS technology in 1990s, the two aspects of wireless devices that have benefited most are cost and form factor. Consequently, demand in the wireless market has grown rapidly, bringing about many new applications and services. In recent years, demand in the wireless market has not been restricted to voice communications, but has also included high-speed data communications such as wireless LAN (802.11a/b/g) and third-generation (3G) cellular communications. Such developments require innovations in design technique to build wireless transceivers with more functionality, while continuing to reduce cost and form factor.

Recent research in wireless transceiver design has focused on full integration of a transceiver using low-cost CMOS technology. This system-on-chip (SoC) integration trend has already been demonstrated in some wireless

systems, such as Bluetooth and wireless LAN [1], [2]. However, in applications such as cellular communications, which require a high-performance transceiver, current handset units still use multiple integrated circuits and discrete components. Figure 1.1 shows photos of two GSM cellular phone circuit boards. The phone on the left was released into the US market in approximately 1995, and the phone on the right in 2003.

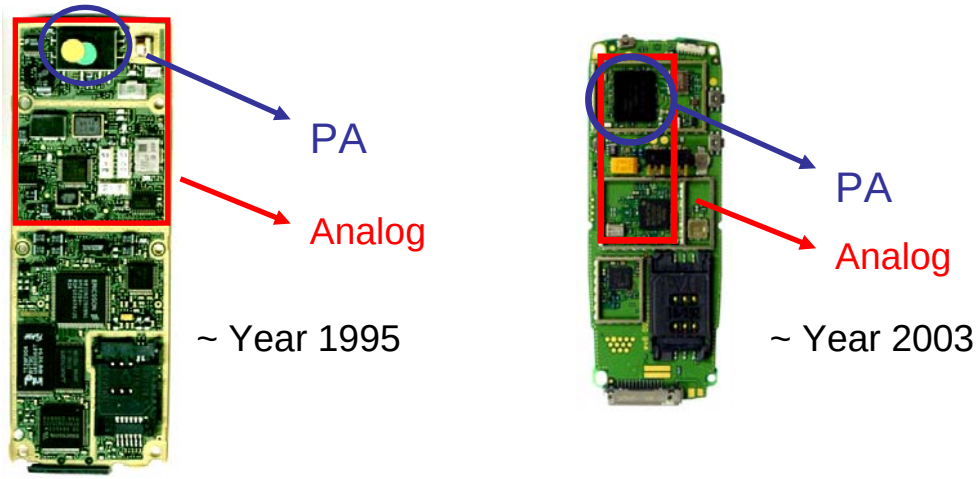


Figure 1.1: Printed circuit boards for GSM cellular phones

After several years of research and development, the number of integrated circuits in a transceiver has lessened and many discrete components have been eliminated. Nonetheless, the power amplifier (PA) has not yet been integrated with the rest of the transceiver. Discrete PAs are usually implemented in a much more costly technology, such as GaAs or LDMOS. A PA in these technologies offers much better performance than its CMOS counterpart but cannot be integrated with the rest of the transceiver.

Much recent research focuses not only on designing a single PA block in CMOS technology, but also on PA enhancement techniques to compensate for shortcomings of the PA itself. The main goal of these enhancement techniques is to improve PA performance beyond what can be achieved by a stand-alone PA. Enhancement techniques have not been widely used in the past since they typically require a significant number of peripheral circuits, which are not easily implemented in technologies commonly used to build a PA. Even though CMOS technology offers a PA with inferior performance, it has an advantage in its ability to integrate and apply one or more enhancement techniques with little overhead cost. This can potentially allow a CMOS PA to be comparable in performance to those built using more expensive technologies. This is especially true for some of the enhancement techniques that are realized in the digital domain, as they will benefit continuously from ongoing technology scaling.

1.2 Power Amplifier in Modern Wireless Applications

In typical wireless transceivers, the PA is the last active building block in the transmitter chain. The foremost function of a PA is to amplify the signal to a level high enough that it can travel through the air to the intended receiver. Often, the PA dominates the power consumption of the entire transceiver. Therefore, improving the PA's power efficiency, defined as the ratio of the output power to the DC power consumption, usually translates to longer battery life in the wireless device.

In some wireless standards, such as GSM, where the modulated signal has constant envelope, a nonlinear PA can be used. A nonlinear PA generally has higher efficiency than a linear PA, as will be explained in Chapter 2. Despite their higher efficiency, the data that these transceivers can transmit within a given bandwidth are low. In order to increase the data rate without increasing the bandwidth, a more spectrally efficient modulation scheme must be used. This makes the modulated signal's envelope non-constant and therefore requires a linear PA to faithfully amplify the signal (Section 2.3). Table 1.1 summarizes specifications of several wireless standards in today market.

Table 1.1: Specifications of several wireless standards

Standard	Freq. Band	Channel spacing	Modulation	Symbol/chip rate	Peak Power
UMTS (FDD)	1.9GHz	5MHz	QPSK	3.84Mcps	2W
802.11a	5GHz	20MHz	2 to 64 QAM	250ksps	800mW
802.11b	2.4GHz	25MHz	CCK	11Mcps	1W
802.11g	2.4GHz	25MHz	2 to 64 QAM	250ksps	1W
Bluetooth	2.4GHz	1MHz	GFSK	1Msps	100mW
GSM	1.8GHz	200kHz	GMSK	270.833kbps	1W
GSM-EDGE	1.8GHz	200kHz	8 PSK	270.833kbps	1W

1.3 Research Goals

The main objective of this research is to design a linear and efficient CMOS PA. This research work can be divided into two significant parts. The first focuses on using several design techniques to improve and optimize the performance of a single CMOS PA. The second focuses on applying

enhancement techniques to further improve the PA performance. The principal enhancement technique used in this research is called the Doherty amplifier; it allows a PA to achieve high efficiency over a wide range of output power. This technique will be discussed further in Chapters 3 and 4.

To demonstrate the design techniques, an experimental prototype was designed to meet Enhanced Data-rate for GSM Evolution (GSM EDGE) specifications. Considered by many as a 2.5G cellular standard, GSM EDGE is an extension of the widely used GSM standard. Rather than using GMSK modulation scheme as in the GSM standard, it adopts a more spectrally efficient modulation scheme ($3\pi/8$ 8-PSK), yielding a higher data rate in the same 200kHz signal bandwidth. This, in turn, requires a linear transmitter to faithfully transmit the signal. The prototype was designed to be a class E1 mobile unit (with 30dBm output power) operating in the DCS1800 band (1710-1785MHz).

The primary contributions of this research are:

- Several PA enhancement techniques were investigated in an attempt to find a suitable technique that can be applied to the CMOS PA problem. These included many efficiency enhancement and linearization techniques. One that was deemed usable is the Doherty amplifier, which is the main topic of this thesis. With the Doherty amplifier, the efficiency at large back-off power can be significantly improved. It can potentially improve peak efficiency as well. Other techniques such as Cartesian feedback or predistortion can be used in conjunction to further improve the efficiency of the Doherty amplifier.

- Analysis of the Doherty amplifier is given and many practical implementation issues are discussed. Doherty amplifier technique is suitable for both linear and nonlinear applications. It can also be extended to have more than two stages and thus achieve a wider high efficiency range.

- The linear class AB CMOS PA was examined. When properly designed, a class AB amplifier can have linearity comparable to that of a class A amplifier, but with significantly higher efficiency. A biasing scheme that allows a class AB amplifier to preserve its linearity across process and temperature variations is discussed.

- A prototype of a Doherty amplifier was designed in a 0.13μm CMOS process with metal-insulator-metal (MIM) capacitors. The prototype targeted GSM EDGE application with over 30dBm of saturated output power. The main amplifier was biased in a class AB region for linearity. The auxiliary amplifier was biased in a class C region for higher overall efficiency. In this prototype, the conventionally used quarter-wavelength transmission line was replaced by a passive lumped-element network to allow full integration. This passive network performs an impedance inversion function, which enables efficient power combining of both amplifiers.

- The prototype was fabricated and its performance measured. It achieves over 31.8dBm of saturated output power, while meeting the GSM spectral mask for all output power levels. The peak phase error was measured to be 1.2° (rms phase error is less than 1°). The measured peak power-added efficiency (PAE) is 36% and stays above 18% over a 10dB range of output power. With an EDGE modulated signal, the prototype achieves a peak output power of 25dBm while still meeting the mask requirement, with a PAE of 13%. The PAE at 12dB back-off power is 6%.

1.4 Thesis Organization

This thesis is organized as follows:

In Chapter 2, fundamental concepts of radio transmitters and RF PAs are presented. The chapter also explains several metrics used for quantifying PA efficiency and linearity. Tradeoffs of different PA classes are discussed. Different PA classes, including A, AB, B, C, D, E, and F are described.

Chapter 3 focuses on two categories of enhancement techniques that can be used to improve PA performance beyond that of a stand-alone PA: linearization and efficiency enhancement. Most of the commonly known techniques are discussed, with their pros and cons.

Chapter 4 reports the research work, which applies the Doherty amplifier technique to the CMOS PA problem. The architecture and detailed operation of this efficiency enhancement technique are discussed.

Chapter 5 is dedicated to a linear stand-alone CMOS PA design, along with several circuit techniques to improve its linearity. Instead of using a class A amplifier, which has poor efficiency, a linear and more efficient class AB amplifier is proposed. A biasing scheme that guarantees good linearity of the class AB amplifier across all process and temperature variations is also described.

In Chapter 6, the prototype that was built to demonstrate the concepts explained in the earlier chapters is explained. The performance measurement results are presented.

Chapter 7 offers conclusions.

Chapter 2

RF Power Amplifier Fundamentals

This chapter covers fundamental knowledge of RF power amplifiers. First, the basic functions of a PA in a radio transmitter are discussed. Then the importance of a PA's efficiency and linearity to the overall performance of a radio transmitter is underlined. To measure these performance aspects, several metrics are introduced. Once the basic functions and metrics of a PA are established, the next section covers conventional PA circuits and operations. Several different PA classes and their schematics are presented. These can be roughly divided into transconductance PAs and switching PAs. Transconductance PAs include classes A, AB, B and C, while classes D and E are of the switching type. Each class has its own merits and drawbacks and is discussed in detail. At the end of this chapter, it will be apparent why the linearity-efficiency tradeoff exists. This tradeoff is very difficult to overcome with conventional design techniques.

2.1 Power Amplifier Basics

Radio transceivers in modern wireless communication systems are almost exclusively digital. In these systems, data that need to be transmitted are generated by a digital baseband circuit and then converted into the analog domain by a digital-to-analog converter (D/A converter). The analog baseband signal is then frequency-translated up to RF by a quadrature modulator, which usually consists of a pair of mixers driven two local oscillators with a 90° phase difference. The signal strength of the modulated RF signal is usually too low to be transmitted, so a PA is needed to amplify the signal to an adequate output power level before it can drive the antenna and be transmitted into the air. Figure 2.1 is a block diagram of a typical radio transmitter. Details of radio transceiver architectures will not be discussed here. Readers are encouraged to refer to [3],[4] for further information.

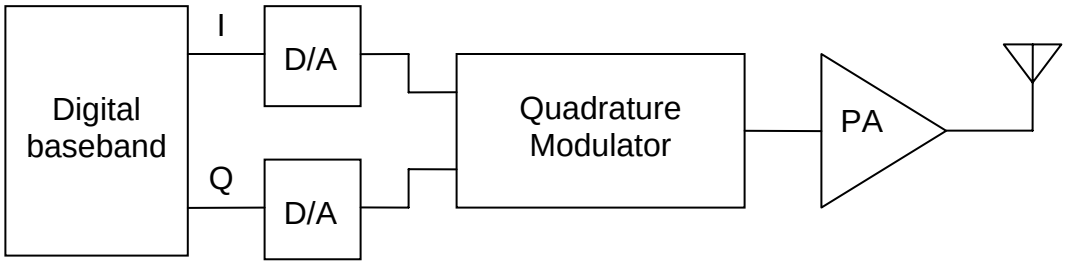


Figure 2.1: Block diagram of a typical radio transmitter

The output power level of a transmitter varies, depending on applications. The major determining factor is the distance that the RF signal has to travel in the air. The output power level of a radio transmitter ranges from the sub-milliwatt

level in short-ranged and low-powered systems through the watt level in cellular handsets to thousand of watts in base stations or satellite communication applications. Apart from using milliwatt (mW) or watt (W) to refer to a power level, it is common to use decibel milliwatts (dBm) or decibel watts (dBW) in high-power systems. Both dBm and dBW are in the logarithmic scale and are referenced to 1mW and 1W, respectively. Throughout this thesis, only dBm is used. P_{dBm} is defined as

$$P_{dBm} = 10 \log_{10}\left(\frac{P}{1mW}\right) \quad (2.1)$$

where P is power in watts. This should not be confused with the definition of dB. dBm and dBW refer to an absolute power level whereas dB is used to show the relative magnitude of two quantities. For example, one can say that the output power of handset A is 30dBm but it is 3dB lower than that of handset B (2x lower when compared to handset B).

Output power levels of several wireless standards are shown in Table 1.1 in Chapter 1.

2.2 Power Efficiency

As previously noted, the main function of a PA is to deliver an appropriate output power level to the load antenna. The delivered output power is obtained by converting the DC power from the power supply into AC signal at the output of the amplifier. Unless the PA has an ideal power conversion ability, some of the DC power does not get converted to the output load but is dissipated and wasted

elsewhere. The metric that is used to measure this conversion ability is efficiency, η, which is defined as

$$\eta = \frac{P_{load}}{P_{DC}} \tag{2.2}$$

where P_{load} is the power delivered to the load and P_{DC} is the DC power drawn from the power supply. For ideal power conversion, η is 1 or 100%. However, a PA's efficiency hardly approaches 100% in reality.

Often, power consumption in a PA dominates the power consumption of the entire transceiver, so it is of paramount importance to design a PA that is as power efficient as possible in order to achieve a power-efficient transmitter. The importance of power efficiency is two-fold. In portable systems where the power supply is a battery, the power efficiency needs to be high in order to maximize battery life. In high-power systems such as base station transmitters, thousands of watts are transmitted and the power loss is usually in the form of heat dissipated in the transmitters. Inefficient transmitters in these applications call for a more sophisticated cooling system, thus increasing the implementation cost. For example, for a PA that has 10,000W output power, an efficiency increase from 40% to 50% equates to 5,000 Watts less heat dissipation (assuming all power loss is through dissipation in the form of heat).

A more technology-specific metric commonly used to measure the power conversion ability of a PA is drain efficiency (if a MOS transistor is used) or collector efficiency (if a bipolar transistor is used). Drain efficiency, η_D, is defined as

$$\eta_D = \frac{P_{load}}{P_{DC,drain}} \quad (2.3)$$

where $P_{DC,drain}$ is the DC power supplied to the amplifier (usually only of the final output stage).

Many PAs also have more than one stage of amplification. The main reason is that concerns over stability impel designs that prevent an amplifying stage from having more than 15-20dB of power gain. In cases where very high output power is desired, a pre-amplifier stage or stages are required. η_D represents the efficiency of only the final output stage. It does not include the amount of DC power consumed in the driving stage or stages. Therefore, with a multistage amplifier, overall efficiency ($\eta_{overall}$) is a more representative number.

$$\eta_{overall} = \frac{P_{load}}{P_{DC,total}} \quad (2.4)$$

where $P_{DC,total}$ is the total DC power consumed in a PA. With a single-stage amplifier, η_D and $\eta_{Overall}$ are the same.

However, there is still one important aspect of the power conversion ability that is not taken into account by either η_D or $\eta_{overall}$, and that is the power required to drive the PA input. Consider a case in which the output stage of a PA is capable of sufficiently driving the output load but still requires a significant amount of input drive power. This implies that the preceding circuit must deliver a large amount of power to drive the input of the output stage. When taking into account the DC power consumption of the driving stage, the overall transmitter efficiency can be significantly lower than the drain efficiency. The metric that is

widely used to include the effect of the input drive power is power added efficiency (PAE), which is defined as

$$PAE = \frac{P_{load} - P_{in}}{P_{DC,total}} \quad (2.5)$$

where P_{in} is the power of the input drive signal.

Of all these efficiency metrics, η_D is the highest number and can sometimes be misleading if misused, as it does not include the power of the driving stage. For a well-designed PA, $\eta_{overall}$ and PAE should be approximately the same and best represent how efficient a PA is.

2.3 Linearity

As described in Chapter 1, many modern wireless communication standards use modulation schemes in which the transmitted signal has a non-constant envelope. The transmitter for these applications must have good linearity in order to preserve the information in both the amplitude (envelope) and phase, and transmit the signal faithfully. Nonlinearities in these transmitters translate to an increase in bit error rate of the received signal, which can corrupt the integrity of a wireless communication link. To understand the need for linearity from a mathematical standpoint, consider a non-constant envelope signal

$$x(t) = A(t)\cos\left(\omega_{RF}\, t + \phi(t)\right) \quad (2.6)$$

The information of the transmitted signal is embedded in A(t) and ϕ(t) in the forms of amplitude, *i.e.*, envelope and phase, respectively. ω_{RF} is the RF carrier

frequency. Assume that nonlinearities in a transmitter can be captured in a power series as follows:

$$y(t) = a_1 x(t) + a_3 x^3(t) + a_5 x^5(t) + \ldots \tag{2.7}$$

Note that even order nonlinearities are not considered here as they do not generate any in-band distortions. With the input signal in Equation 2.6, and only third- and fifth-order nonlinearities included, y(t) is found to be:

$$\begin{aligned} y(t) = & \left[a_1 A(t) + \frac{3}{4} a_3 A^3(t) + \frac{5}{8} a_5 A^5(t) \right] \cos\left(\omega_{RF} t + \phi(t)\right) \\ & + \left[\frac{1}{4} a_3 A^3(t) + \frac{5}{16} a_5 A^5(t) \right] \cos\left(3\omega_{RF} t + 3\phi(t)\right) \\ & + \left[\frac{1}{16} a_5 A^5(t) \right] \cos\left(5\omega_{RF} t + 5\phi(t)\right) \end{aligned} \tag{2.8}$$

From Equation 2.8, the first term, which is the in-band component, has its envelope distorted by the nonlinearities of the transmitter transfer characteristic. Therefore, all odd-order nonlinearities must be minimized to avoid losing the information in the envelope of the transmitted signal. Note that the phase information of the in-band component is unharmed by the transmitter nonlinearities. This is why a nonlinear transmitter can be used with a constant envelope signal. However, this statement is true only if the transmitter nonlinearities can be written in a power series format. In cases where the transmitter has a strong memory effect, a Volterra series must be used instead and the information contained in the phase of the signal may be distorted. One example of this situation is when the frequency response for the band of interest

is not flat, and therefore the output spectrum is distorted regardless of whether the signal has a constant or a non-constant envelope.

Most circuits in a transmitter only have to deal with small signals. Linearity in these circuits is important but is not as critical as in PA where the output signal has large voltage and current swings. The voltage swing at the output of a PA is usually designed to be as large as possible in order to achieve high efficiency. Linearity in a PA can be quantified in several ways and will be discussed next in detail.

2.3.1 AM-to-AM and AM-to-PM Characteristics

AM-to-AM (amplitude modulation to amplitude modulation) is the relationship between the amplitude of the input signal and the amplitude of the output signal. AM-to-PM (amplitude modulation to phase modulation) is the relationship between the amplitude of the input signal and the phase of the output signal. Both metrics use sinusoidal input signal and only the fundamental component of the output is measured. Linearity of the AM-to-AM characteristic can sometimes be equivalently characterized as P_{1dB}, which is the output power level when the gain is compressed by 1dB compared to the small signal gain. AM-to-PM shows how phase changes with different input amplitudes. These two metrics are widely used to measure crudely how linear a PA is. They are easy to measure or simulate since only sinusoidal input signal is involved. The downside is that they do not capture some of the memory effects, since only the output amplitude or phase at steady state is measured, whereas for non-constant

envelope signal the amplitude of the RF signal can vary rapidly. Figure 2.2 shows the typical AM-to-AM and AM-to-PM characteristics of a linear PA.

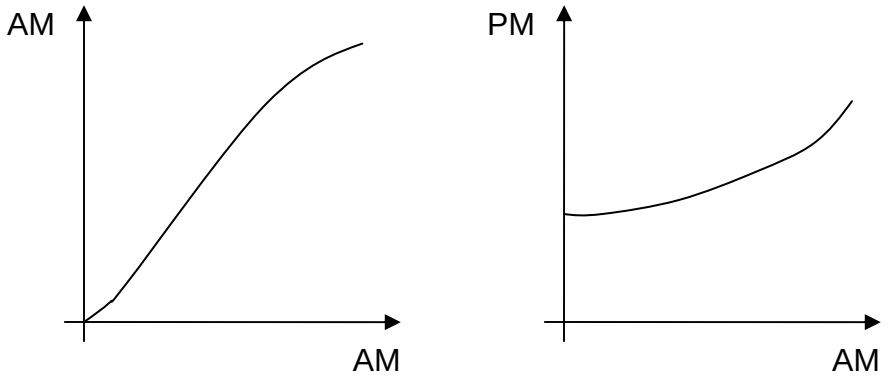


Figure 2.2: Typical AM-to-AM and AM-to-PM characteristics of a linear PA

2.3.2 Spectral Mask

In a narrow-band wireless communication system, the spectrum of the transmitted signal should ideally be strictly confined within the specified channel bandwidth. However, in reality it is not possible to have no power leakage into adjacent bands, since a band-limited signal corresponds to a signal that is infinitely long in the time domain. Therefore, a proper band-limiting filter, sometimes called a pulse-shaping filter, must be used (usually pre-specified by the wireless standard). A good pulse-shaping filter should allow very little power outside the transmit channel. When a nonlinear transmitter is used, out-of-band intermodulations can be generated, causing more power to leak into adjacent bands. Without proper control, this power leakage can corrupt the signal that resides in the neighboring channel. One way to specify the linearity requirement of a transmitter is by defining a spectral mask. A spectral mask gives an upper

bound to how much power can be transmitted at frequencies close to the RF carrier. Figure 2.3 shows the spectral mask for GSM-EDGE handsets in the PCS band.

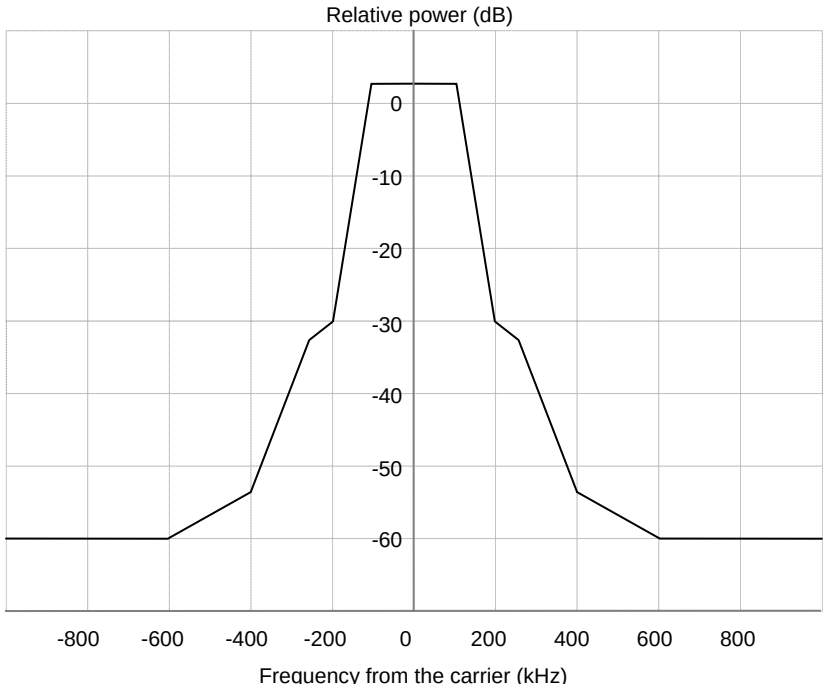


Figure 2.3: Spectral mask for GSM-EDGE handsets in the PCS band

2.3.3 Adjacent Channel Leakage Ratio

Another metric to specify PA or transmitter linearity is the adjacent channel leakage ratio (ACLR), also known as adjacent channel power ratio (ACPR). ACLR is the ratio of the total power in a band to the power in the adjacent band of interest. The ACLR requirement for each adjacent channel varies based on how far the channel is from the RF carrier. As an example, for any UMTS (FDD) handset, ACLR must be less than -33dB for the first adjacent channel (5MHz away for the RF carrier) and -43dB for the second adjacent

channel (10MHz away for the RF carrier) [5]. Some standards such as GSM do not explicitly specify the ACLR requirement.

2.3.4 Error Vector Magnitude

The two previous linearity metrics measure out-of-band nonlinearities. They are used to guarantee that the power leaking into adjacent bands is not excessive, but they do not ensure the integrity of the transmitted data that reside within the original channel bandwidth. The most common measure for insuring good in-band data transmission is the error vector magnitude (EVM). To measure the EVM of a PA or a transmitter, a set of random bit streams is generated in the digital baseband and then modulated according to the modulation scheme specified by the wireless standard. This baseband signal is then frequency-translated to RF by ideal mixers (actual baseband circuits and mixers are used in case of transmitter EVM measurement) and then amplified by the PA under test. The output signal is received and demodulated by an ideal receiver. The received signal constellation that is obtained after baseband detection is compared with the initial constellation on the transmit side. EVM, measured on the I-Q plane, is the ratio of the error vector length to the distance from the origin of a transmit data point, as shown in Figure 2.4a. Figure 2.4b shows an ideal constellation of a quadrature phase shift keying (QPSK) signal and Figure 2.4c shows the received signal constellation when a typical linear PA is used. As an example of EVM specifications, the GSM standard requires that a handset have an RMS EVM of less than 9% under normal conditions.

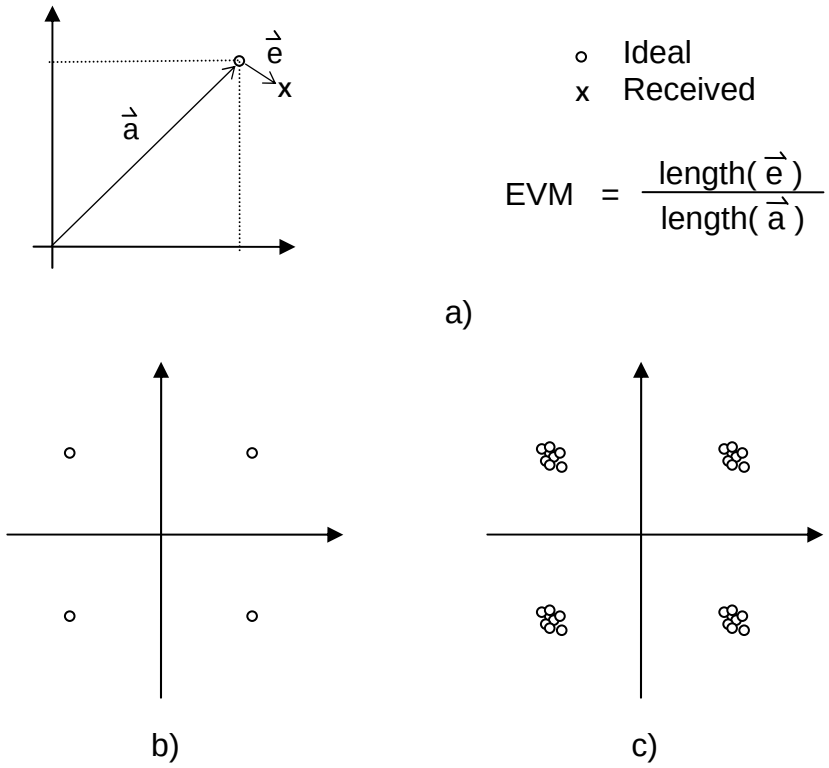


Figure 2.4: Error vector magnitude (EVM)

2.4 A Typical PA Circuit

A typical PA circuit consists of an active device and a load network. Active devices commonly used in PAs are the hetero-junction bipolar transistor (HBT), the bipolar junction transistor (BJT), and the field effect transistor (FET). They are used to provide signal amplification. A load network typically consists of a load resistance, a lossless (or low loss) impedance transformation network, and a bandpass filter.

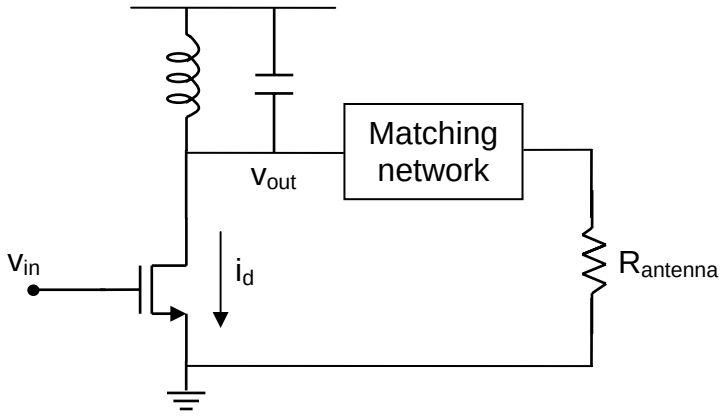


Figure 2.5: Typical PA circuit

The load resistance is usually the characteristic impedance presented by the antenna. Typical values are 50Ω and 75Ω. In Figure 2.5, an LC resonator at the output of the PA acts as a bandpass filter, which filters out unwanted out-of-band signals. The capacitor in the LC tank is mainly from the output capacitance of the active device. This requires an inductor to resonate, yielding a bandpass filter as a by-product. Even though the LC tank is a second-order filter, its rolloff is quite gentle due to the low-Q nature of the output node. Additional filters, such as a SAW filter, may be used to further filter out the unwanted signals.

The need for an impedance transformation network at the PA output can be demonstrated with the following example. Consider a PA with a maximum voltage swing of 3V. Assuming a 50Ω load impedance from the antenna, the maximum output power is

$$P_{out} = \frac{1}{2}\left(\frac{3^2}{50}\right) = 90\,mW \qquad (2.6)$$

In order for this amplifier to output 1W, a load resistance of 4.5Ω is needed. Therefore, a step-down impedance transformation network is required. This network should ideally be lossless; otherwise it will directly degrade the PA's efficiency.

2.5 Power Amplifier Classes

Power amplifiers can be divided into two categories based on how the active device behaves. The first is the transconductance PA, which makes use of the active device as a transconductor. In this category, the active device acts as a voltage-controlled current source, meaning its output current is controlled by its input voltage. The second category is the switching PA. PAs in this category use the active device as a switch to modulate the output voltage or current. Transconductance PAs include classes A, AB, B, and C, while classes D and E are switching PAs. Class F is commonly understood to be a switching PA, but in fact, it can also be a transconductance PA, depending on how hard the active device is driven.

2.5.1 The Class A Amplifier

In a class A amplifier, the transistor is biased so that it is turned on at all times. In terms of conduction angle, which is how long the device is turned on in one RF cycle, class A has a conduction angle of 360°. Typical waveforms of a class A amplifier are shown in Figure 2.6.

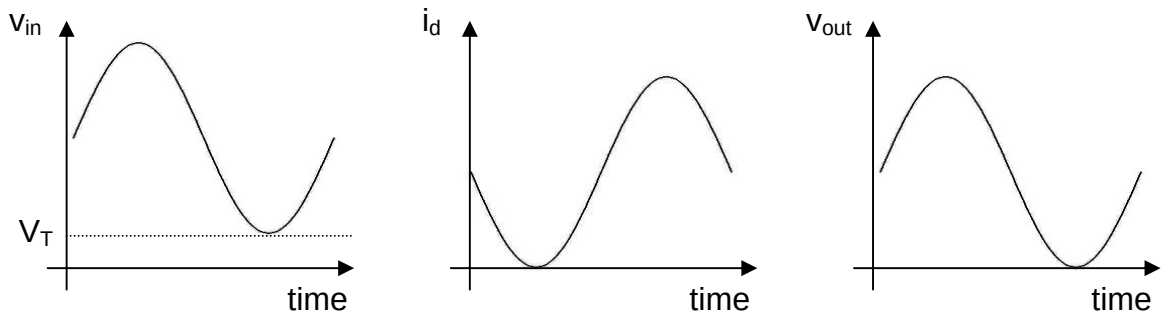


Figure 2.6: Class A amplifier waveforms

In Figure 2.6, it is assumed that the transistor has linear transconductance, $i_d = k(v_{in} - V_T)$. This in turn, gives a linear relationship between the fundamental component of the output voltage and the input voltage. Since the device is turned on all the time, no nonlinearities due to signal clipping are introduced. For this reason, a class A amplifier is often considered to be the most linear of all. However, if the transconductance of the active device is not linear, this transfer characteristic of the fundamental component will most likely be nonlinear, but it is still better than other amplifier classes in general. This is because there is no distortion caused by the active device's switching off. The maximum achievable efficiency of class A amplifiers is 50% (for a linear transconductance device). In reality, the obtainable efficiency rarely exceeds 30-40% due to nonidealities in the transistor and finite Q of the passive components.

When a high level of linearity is required, the output swing has to back off significantly from the maximum swing to avoid excessively large signal distortions. In some cases, the efficiency may even need to be allowed to fall below 10% [6] in order to meet the desired linearity requirement.

It is particularly interesting to explore the behavior of a class A amplifier when the active device transconductance obeys the square law, as in long-channel CMOS transistors. With square law transconductance (ignoring the threshold voltage),

$$i_d = k.v_{in}^{\ 2} \tag{2.7}$$

Let $$v_{in} = V_o(1+\sin \omega t) \tag{2.8}$$

This gives $$i_d = k.V_o^{\ 2}(1.5+2\sin \omega t + 0.5\cos 2\omega t) \tag{2.9}$$

Waveforms of a class A amplifier with square law transconductance are shown in Figure 2.7. Even though the drain current has the second harmonic component, the linearity of the fundamental component is preserved. Besides, since the drain current waveform has a flatter trough due to the presence of the second harmonic component, the DC current is less than in the linear transconductance case. In this case, the peak efficiency was found to be 66%.

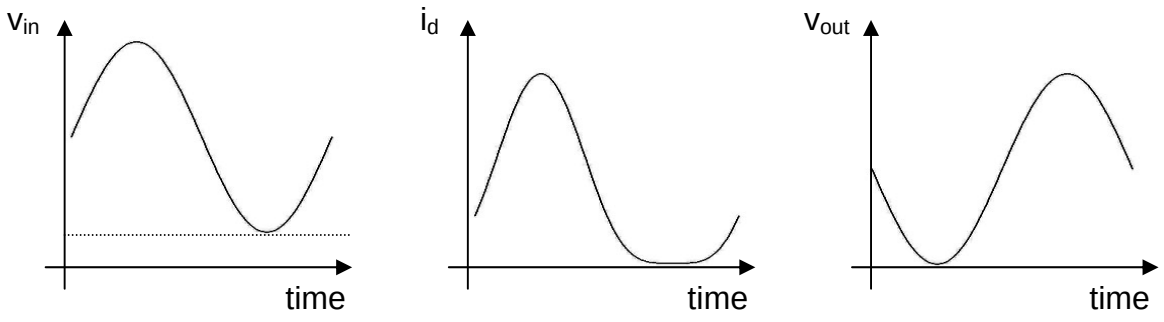


Figure 2.7: Waveforms of a class A amplifier with square law transconductance

It is also noteworthy that the amount of DC power drawn from the power supply of a class A amplifier is constant. This power is either converted and then delivered to the load or dissipated in the amplifier itself. As a result, the heat dissipation in the active device is the highest when the amplifier does not supply any output power. As the amplifier supplies more power, the device temperature decreases (assuming all power loss is through dissipation in form of heat).

2.5.2 Class AB, B, and C Amplifiers

In class AB, B, and C amplifiers, the transistor is biased such that it is turned off during some part of the cycle. Class AB has a conduction angle between 180° and 360°. Class B and C have exactly 180° and less than 180°, respectively. Assuming linear transconductance, a class B amplifier can achieve 78.5% efficiency and this can potentially reach 100% for a class C.

Figure 2.8 shows the waveforms of a class C amplifier.

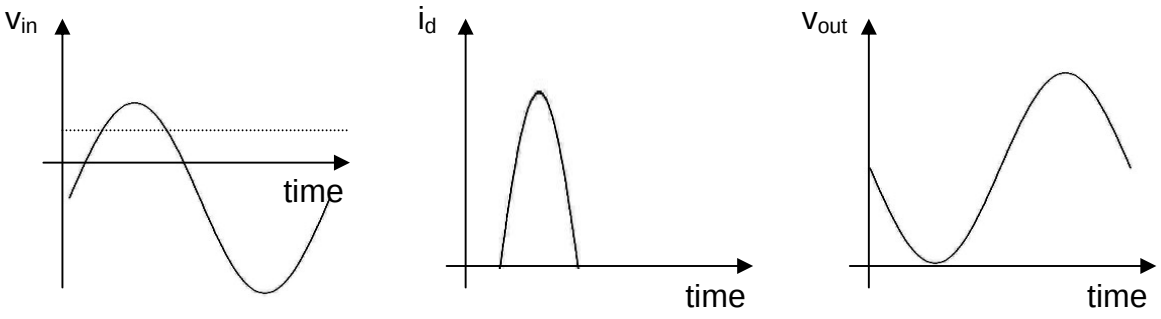


Figure 2.8: Waveforms of a class C amplifier

Even though a class C amplifier can potentially achieve 100% efficiency, in practice it is very difficult to realize such high efficiency with any meaningful output power because of the diminishing conduction angle. In other words, as the conduction angle approaches zero, the fundamental component of the output current also approaches zero. For a fixed sinusoidal input amplitude, Figure 2.9 shows a plot of the transistor sizes needed to obtain the same fundamental output current, versus conduction angle, normalized to the class A case (a transistor with linear transconductance is assumed). It can be seen that as the conduction angle approaches zero, an infinitely large transistor is needed to get the same output current as in class A. Apart from the reduction in the output current, the input drive power also increases as the conduction angle decreases (since the device size keeps getting bigger). This requires more power from the driving stage, causing the overall efficiency to drop. This is one of the reasons why a deep class C amplifier with high efficiency is hard to engineer and often is not practical in reality. To illustrate the effect of the input power drive on efficiency, the PAE of PAs with different conduction angles is plotted in Figure 2.10 (the dotted line). This plot was obtained by scaling the transistor size according to Figure 2.9 in order to achieve the same output power with a fixed input drive level. It is also assumed that the power gain of the class A amplifier is 20dB and the input impedance of the amplifier has constant Q regardless of conduction angle, so that the input power scales with transistor size. The solid line in Figure 2.10 was obtained by assuming zero input drive power and is shown here for comparison.

Another approach to attaining high efficiency is to implement the active device as a switch. Switching amplifiers can also potentially be 100% efficient, like a class C amplifier, but they are proven to be more practical to realize.

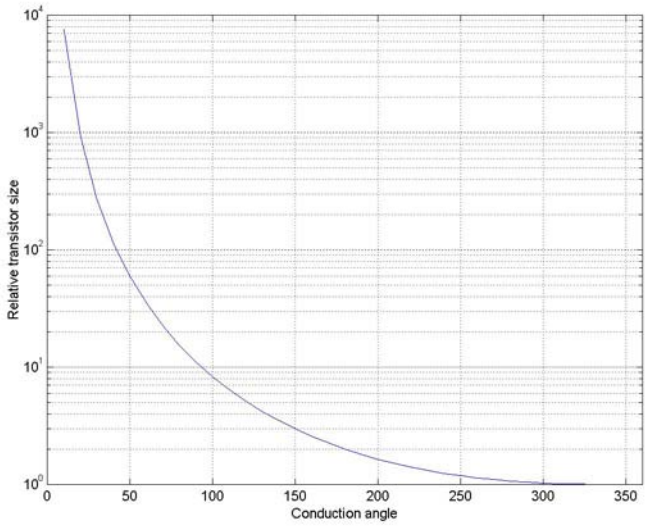


Figure 2.9: Transistor sizing for different PA classes

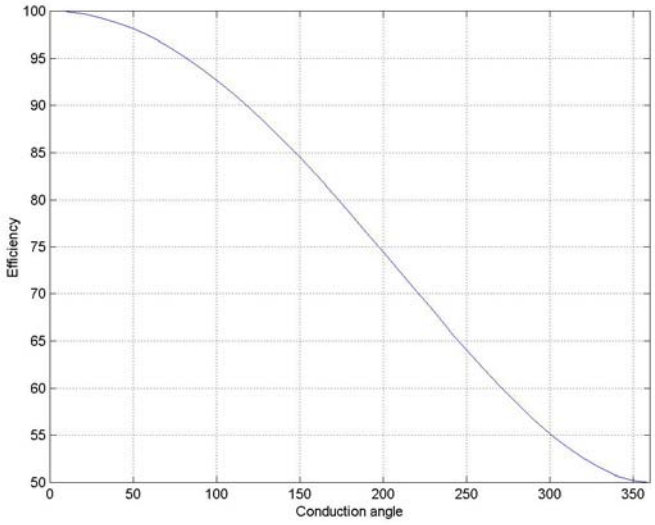


Figure 2.10: Efficiency versus conduction angle

2.5.3 The Class D Amplifier

Figure 2.11 shows a diagram of a class D amplifier. The active device is assumed to behave as an ideal single-pole, double-throw (SPDT) switch. The series LC tank forces the load current, I_o, to be sinusoidal. The waveforms are shown in Figure 2.12. It can be seen that the switch current and switch voltage do not have any overlapping non-zero instance, hence there is no loss in the switch. Assuming all reactive components have infinite Q, 100% efficiency can be achieved. In practice, the presence of parasitic capacitance of the active device prevents V_{sw} from changing abruptly, causing the voltage and current waveforms to overlap. This results in power loss in the switch, and therefore drives efficiency below 100%.

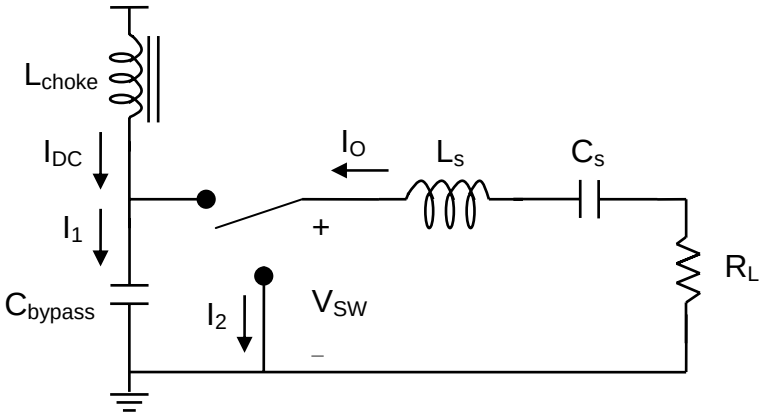


Figure 2.11: Class D amplifier schematic diagram

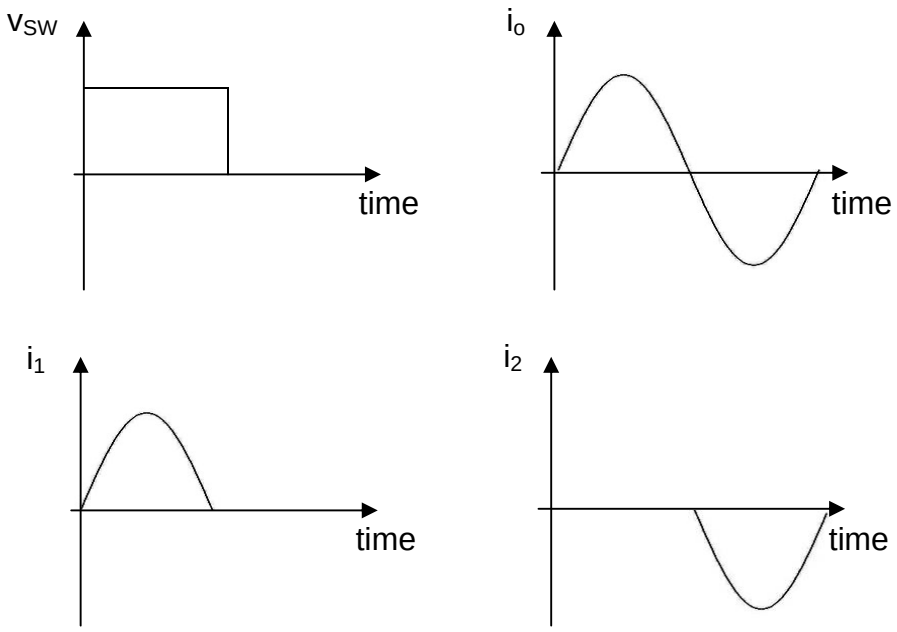


Figure 2.12: Waveforms of a class D amplifier

2.5.3 The Class E Amplifier

A Class E amplifier is also a switching amplifier. It has the potential to be more efficient than any other amplifier classes since it can accommodate some of the switch non-idealities. The only drawback of a class E amplifier is that its peak voltage can be a few times higher than that of its class D counterpart, causing a potential threat to the amplifier's reliability. Figures 2.13 and 2.14 show a schematic diagram and the waveforms of a class E amplifier. Unlike the class D amplifier, here the output parasitic capacitance is absorbed into the output network.

For Figure 2.14, it was assumed that the switch opens from θ_1 to θ_2. Outside this interval, V_{sw} is forced to be zero by the switch's turning on. During

this interval, V_{sw} can be found by integrating the capacitor current waveform, resulting in V_{sw} shown in 2.14d). At θ_2, V_{sw} is certain to return to zero since there can be no charge accumulation from RF cycle to cycle during steady state. The peak efficiency of a class E amplifier is also 100%, since there is no loss in the switch. Details of designing a class E amplifier can be found in [7], [8], and [9].

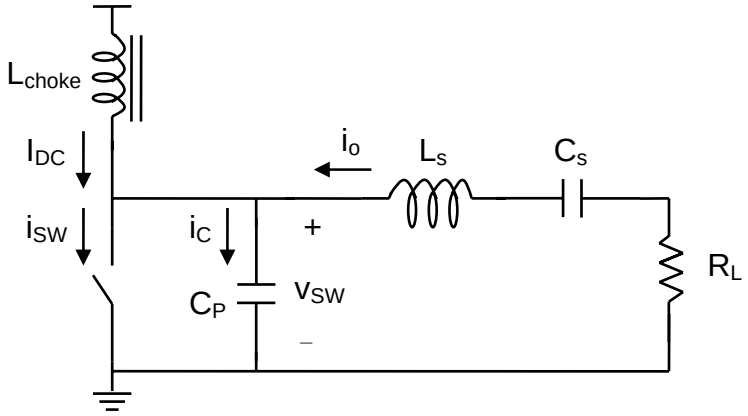


Figure 2.13: Class E amplifier schematic diagram

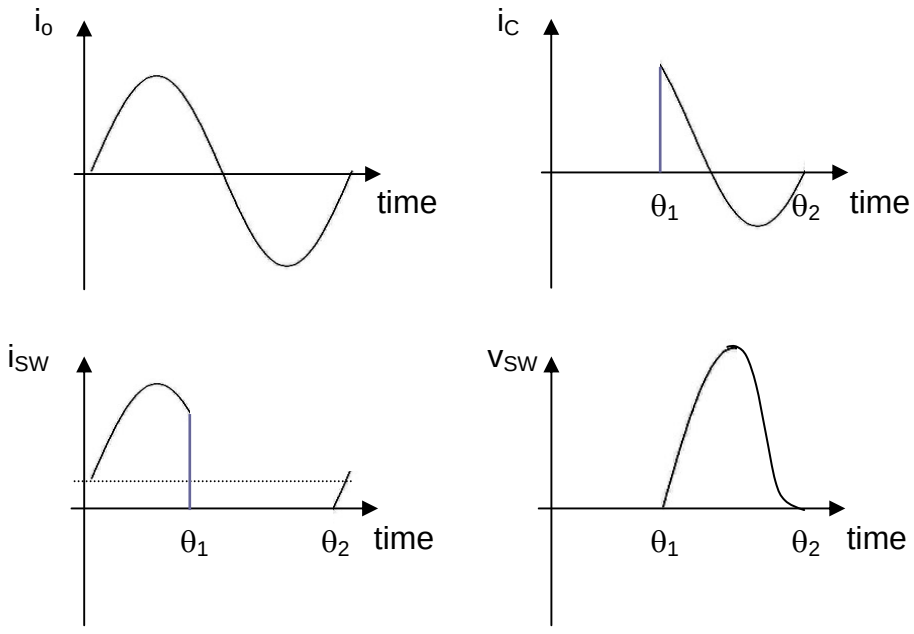


Figure 2.14: Waveforms of a class E amplifier

2.5.4 The Class F Amplifier

The chief concept behind the class F amplifier is to use proper harmonic terminations to square up the output voltage waveform, V_{sw}, such that loss in the active device is minimized. Consider Figure 2.15, in which a parallel LC tank is put in series with the load. This tank is tuned to resonate at the third harmonic and is assumed to have infinite Q such that it does not affect the operation at the fundamental frequency. By having proper magnitude and phase of the third harmonic in the transistor current, V_{sw} starts to resemble a square wave as illustrated in Figure 2.16. By choosing harmonic terminations such that all of the odd harmonics are open and even harmonics are short, a square V_{sw} can potentially be achieved, thus eliminating loss incurred in the switch. Two quarter-wavelength transmission lines can be used to achieve this scenario. This can be done by tuning one of them to the second harmonic and another to the third harmonic.

In order to get harmonic components in the amplifier output voltage swing, these harmonics must be present in the active device's output current. In general, both switching PA and transconductance PA contain a certain amount of harmonic components in their output current (even though much less than in a transconductance PA), so they both can be turned into a class F amplifier with proper load terminations. In practice, due to the finite Q of passive devices, harmonic terminations can potentially be a hindrance rather than a benefit,

especially if on-chip passive devices are used. This is one of the reasons why class F has not yet been widely adopted.

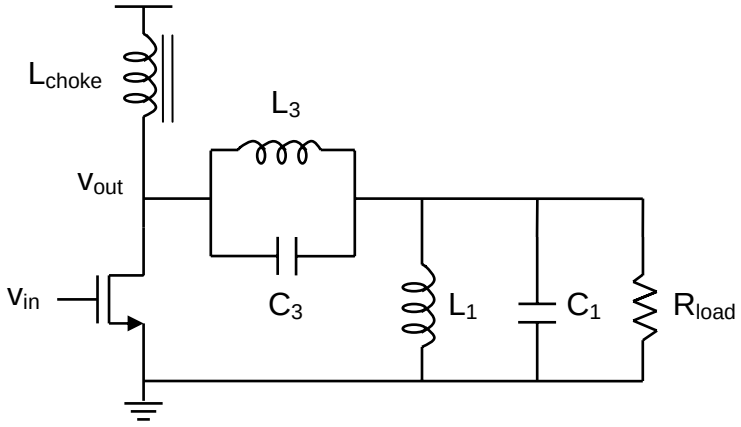


Figure 2.15: Class F amplifier schematic diagram

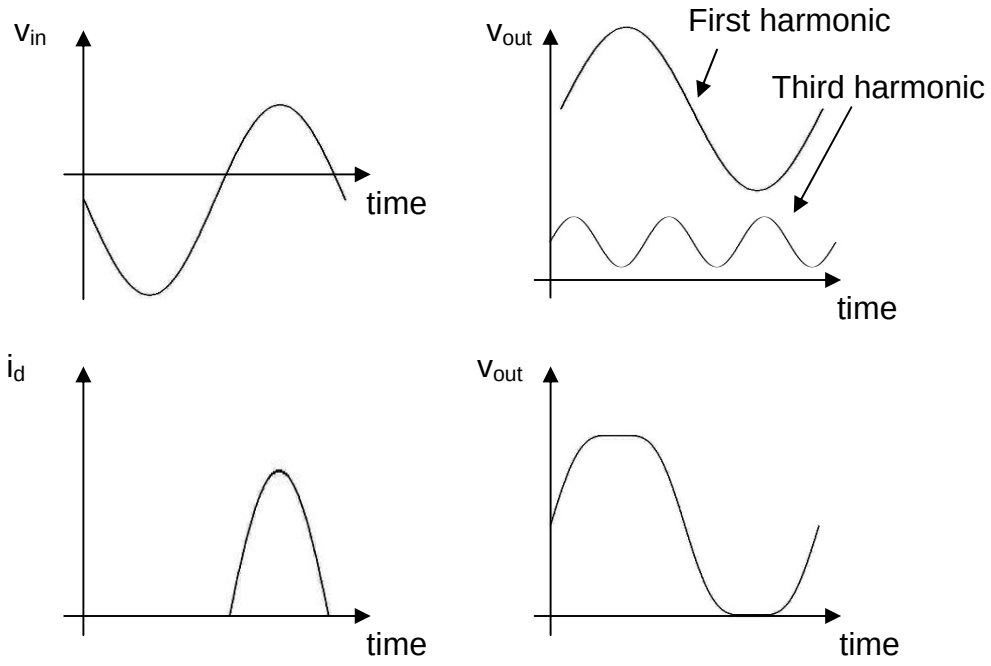


Figure 2.16: Waveforms of a class F amplifier with a third harmonic trap

In order for a switching amplifier to achieve high efficiency, the active device must behave as closely to an ideal switch as possible. In other words, the active device's region of operation should only toggle between cutoff and triode region for a MOS transistor (or saturation region for a bipolar transistor). This can be achieved by allowing the input drive amplitude to be as high as the previous amplifier stage can provide. Under this condition, the amplifier output amplitude does not depend on the amplitude of the input signal (it may depend on the duty cycle of the input signal, though), so there is no linearity in its AM-AM characteristic. This behavior is different for a transconductance PA, where the output amplitude can be controlled by the amplifier input amplitude. It was shown earlier that a class A amplifier can theoretically be perfectly linear but has poor efficiency. By contrast, a class C amplifier has higher efficiency but poorer linearity (though still better than that of switching PAs). This underscores the existence of a linearity-efficiency tradeoff in conventional PA design and inspires our search for PA enhancement techniques to overcome this tradeoff.

Chapter 3

Power Amplifier Enhancement Techniques

In this chapter, several PA enhancement techniques are presented. Their main purpose is to improve the PA performance beyond that of a stand-alone amplifier. As outlined in the previous chapter, there is a fundamental tradeoff between linearity and efficiency in conventional design methods. Enhancement techniques aim to defeat this tradeoff barrier. PA enhancement techniques fall into two categories: efficiency enhancement and linearization. Each has its pros and cons, which will be discussed in detail in sections 3.2 and 3.3. More than one of these techniques can be used simultaneously to further improve PA performance. This is particularly important in CMOS PA design, as they can be used to compensate for shortcomings associated with the technology itself.

3.1 PA Enhancement Techniques

As mentioned in Chapter 2, both efficiency and linearity are of vital importance in PA design for a linear transmitter. However, the linearity-efficiency tradeoff in conventional PA design techniques makes the design of a PA that is both linear and efficient difficult. This challenge is even more severe in CMOS technology where designing either a linear PA or an efficient PA is already challenging in itself. Hence the need for enhancement techniques.

PA efficiency enhancement attempts to improve the efficiency of a linear but inefficient PA. Linearization takes the opposite direction by attempting to improve the linearity a nonlinear but efficient PA.

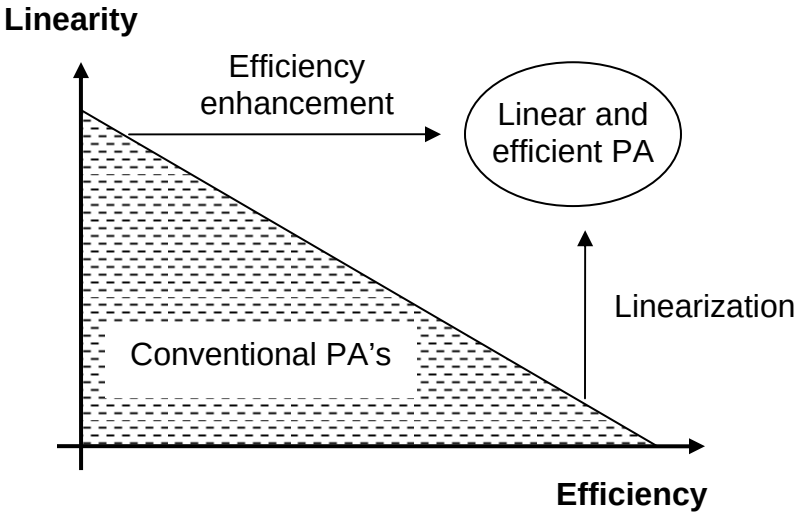


Figure 3.1: Linearity-efficiency tradeoff

The reason that PA enhancement techniques have not yet been widely implemented is because high-end PAs are currently built in expensive

technologies such as GaAs HBT or LDMOS, and these technologies have a much better linearity-efficiency tradeoff characteristic than CMOS. A stand-alone PA in these technologies can perform at a reasonable level and therefore does not require enhancement. Besides, these technologies do not lend themselves well to high integration. As a result, designing peripheral circuits to implement enhancement techniques can be a challenging task.

As attempts to build a single-chip radio increase, integrating a CMOS PA with the rest of the transceiver is an inevitable step. Since CMOS PAs generally have much worse performance than those in GaAs HBT or LDMOS, designing a CMOS PA without enhancement is a mere brute force approach and does not provide any competitive advantage. We will look next at some of these efficiency enhancement and linearization techniques.

3.2 Efficiency Enhancement Techniques

As noted, these are techniques that can be used to enhance the efficiency of a linear but rather inefficient PA. There are two aspects of efficiency to enhance: back-off efficiency and peak efficiency. In most conventional PA configurations, peak efficiency is obtained only when the PA output power is at maximum. The efficiency degrades dramatically once the output power drops below its peak. Most of the time, PA output needs to be much lower than the peak, so its overall efficiency is much less than peak efficiency. Improving the PA efficiency at back-off power can potentially improve the overall efficiency significantly. The first two techniques discussed in this section, the Doherty

amplifier and power supply variation, directly address this problem. They allow a PA to achieve good efficiency over a wide range of output power, yielding better overall efficiency. A third technique, bias adaptation, also improves the efficiency at back-off power. The bias adaptation technique attempts to adjust the bias points of a PA to ensure that the bias current in the active device at each power level is no more than is necessary for optimal efficiency. While this technique is less effective than the first two, it offers the advantage of simplicity in implementation.

Most of the known efficiency enhancement techniques only try to improve efficiency at the back-off power level without affecting peak efficiency. However under some circumstances, the Doherty amplifier technique can be used to improve the peak efficiency as well. This technique will be introduced briefly next and then discussed at length in Chapter 4. Note that all the efficiency enhancement techniques presented in this chapter can also be applied to a nonlinear amplifier to improve its efficiency.

3.2.1 The Doherty Amplifier

The Doherty amplifier technique was invented by William H. Doherty and first published in the Proceedings of Institute of Radio Engineers in September, 1936 [10]. The first commercial transmitter using this technique was the 50kW TV transmitter for WHAS in Louisville, Kentucky [11]. The main objective of this technique is to allow a PA to achieve high efficiency over a wide range of output power.

One common characteristic of conventional PAs is that peak efficiency can be obtained only at peak output power. The efficiency then drops significantly as the output power decreases. For a typical amplifier, peak efficiency can be achieved under the condition:

$$\left.\begin{aligned} P_{out} &= \frac{1}{2}\left(\frac{V_{\max}{}^2}{R_{load}}\right) \\ V_{\max} &= I_{\max} R_{load} \end{aligned}\right\} \qquad (3.1)$$

In general, V_{max} is fixed by an external power supply voltage and R_{load} is fixed by the antenna impedance and the transformation ratio of the output-matching network. Consequently, the peak efficiency can be obtained at only one output power level. Figure 3.2 shows plots of efficiency versus output power of ideal class A, B, and C amplifiers. In this Figure, it is assumed that the active device in use has a linear transconductor, and both the power supply voltage and the load resistance are fixed. At 10dB power back-off, the efficiency of a class A amplifier is reduced by a factor of 10, and for class B and class C amplifiers by a factor of approximately 3.

In order to satisfy the maximum efficiency condition (Equation 3.1) at a different output power level, either V_{max} or R_{load} (or both) must be allowed to vary. V_{max} can be varied by means of a DC-DC converter, as will be discussed in Section 3.2.2. An alternative is to change R_{load}. For a given output power level, if R_{load} is changed in such a way that the output swing is still kept at V_{max}, it is possible to obtain peak efficiency at a power level lower than peak. One way to attempt this is to use the Doherty amplifier.

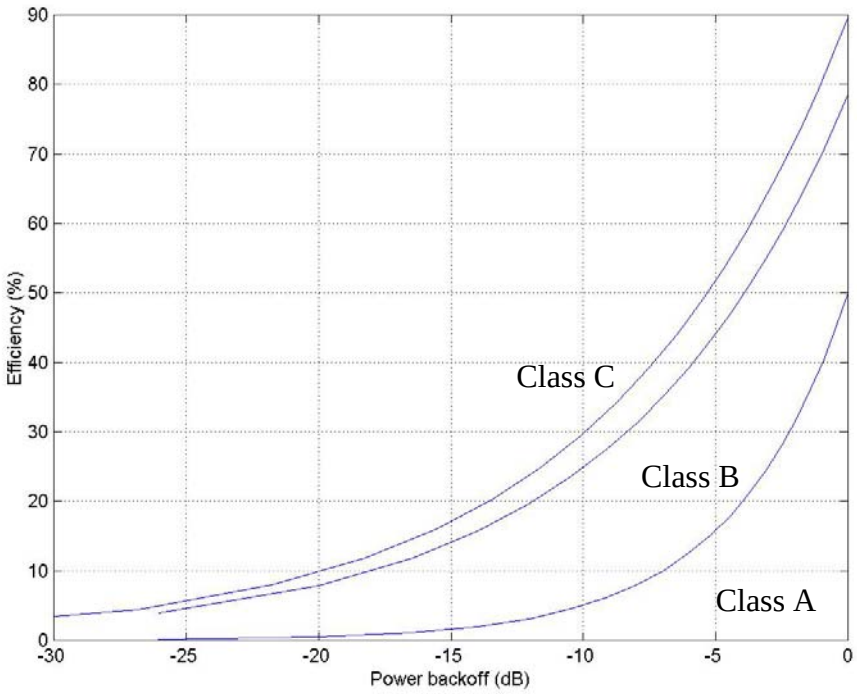


Figure 3.2: Efficiency of conventional PAs

In this technique, two amplifiers are used—main and auxiliary. During low-output power operation, only the main amplifier is turned on. Once the output voltage of the main amplifier reaches the maximum, the first peak in efficiency is obtained. If more output power is desired, the auxiliary amplifier is also turned on. By using a unique way of combining the output power from both amplifiers, the efficiency of the main amplifier is kept at the maximum value all the time during high output power operation.

The key concept of the Doherty amplifier is the technique used to efficiently combine the power from both amplifiers. The two amplifier outputs cannot simply be tied together; true, their output currents (assuming both amplifiers have current source output) will sum, but higher output swing is then

required, which does not improve the efficiency. A microwave power combiner is not desirable since it is not efficient. Instead, the Doherty amplifier technique uses a passive impedance inverter as a combiner. The details of this technique will be discussed further in Chapter 4.

3.2.2 Power Supply Variation

Another way to achieve high efficiency over a wide range of output power is by reducing V_{max} as the output power decreases. This can be done by means of a DC-DC converter. The DC voltage coming out of the DC-DC converter has to be large enough to ensure that there is no signal clipping, otherwise a significant amount of distortion can be generated. The DC-DC converter can use the power control information from the DSP or a power detector to set its output voltage. That voltage is usually changed at a rate much slower than the envelope of the signal itself, so its bandwidth does not need to be large. This allows the DC-DC converter to be very efficient and does not degrade the overall efficiency.

This power supply variation technique can be used to enhance the overall efficiency of both linear and nonlinear PAs. The most important aspect of this technique is in the design of an efficient DC-DC converter. Note that this technique is similar to the Envelope Elimination and Restoration (EER) technique, which will be discussed in Section 3.3.5. The difference is that this power supply variation technique tries to reduce the supply voltage without causing any signal clipping. It still relies on the linearity of the amplifier itself, whereas EER uses a

nonlinear amplifier with a DC-DC converter to vary the supply voltage so as to follow the envelope of the desired signal.

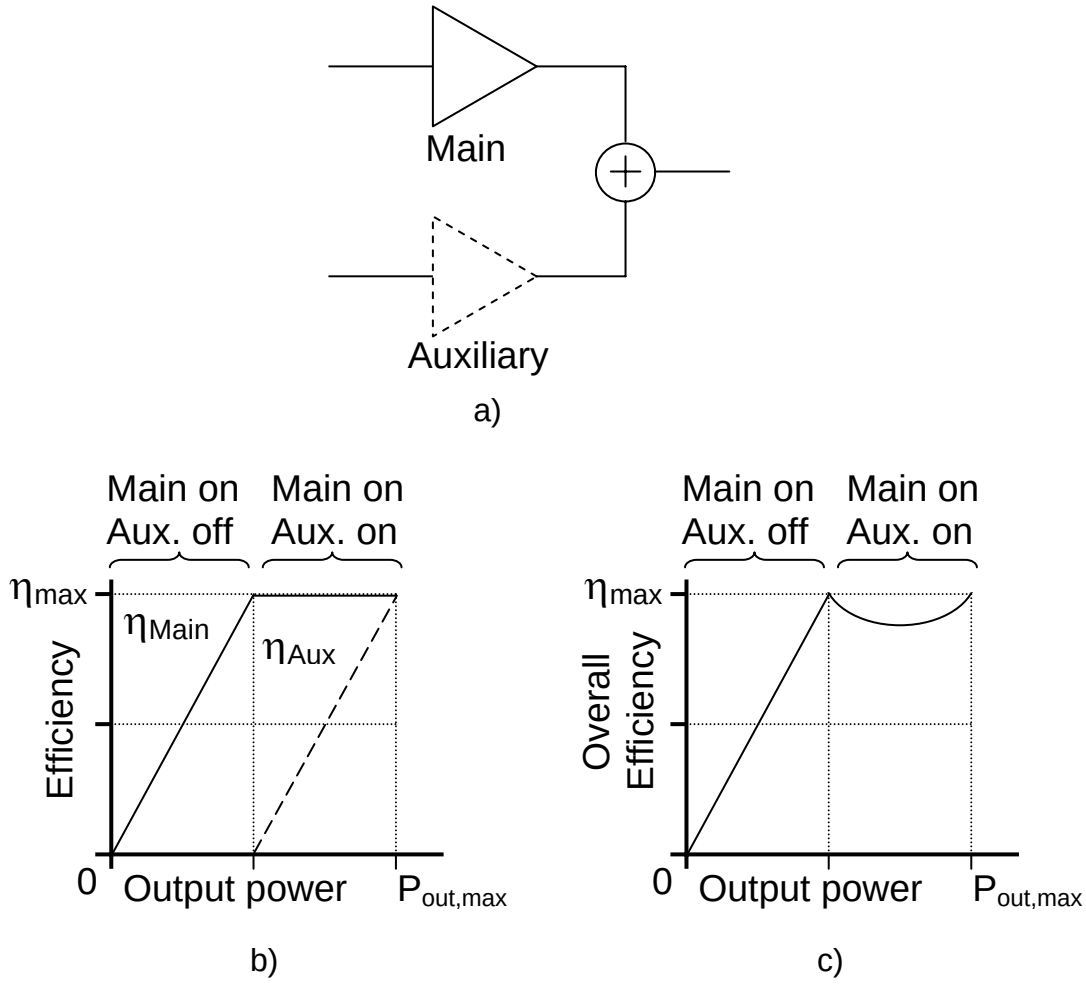


Figure 3.3: a) Conceptual diagram, b) Amplifier efficiency, c) Overall efficiency

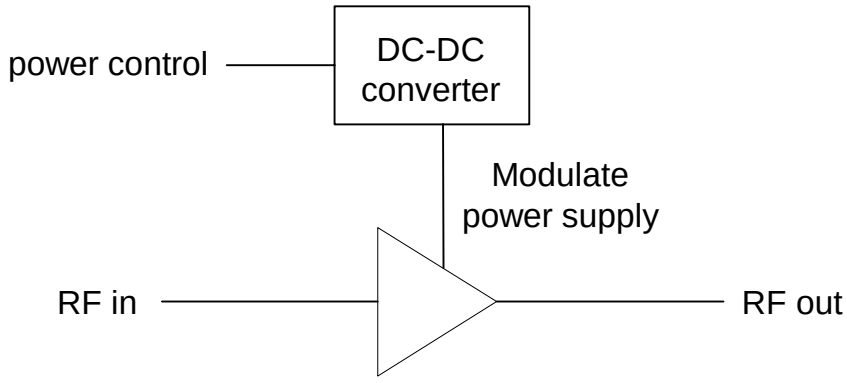


Figure 3.4: Power supply variation technique

3.2.3 Bias Adaptation

Another technique that can mildly improve efficiency is bias adaptation. In a class A amplifier with a fixed bias current, the DC power consumption is constant regardless of the output power, so as the output power decreases from the maximum level, the efficiency also drops linearly. However, if the bias current is adjusted according to the output power level so that the amplifier is at the edge of class AB all the time, the efficiency would drop much more slowly. This concept can also be applied to other amplifier classes to improve the efficiency versus output power characteristic. For a class AB amplifier, bias adaptation may also play an important role in preserving the linearity of the PA.

In order to adjust the bias of the PA, the output power must be detected by a power detector. A look-up table and some digital control circuits are needed to adjust the appropriate bias to the amplifier. An alternative is to use the power control signaling from the baseband or the base station.

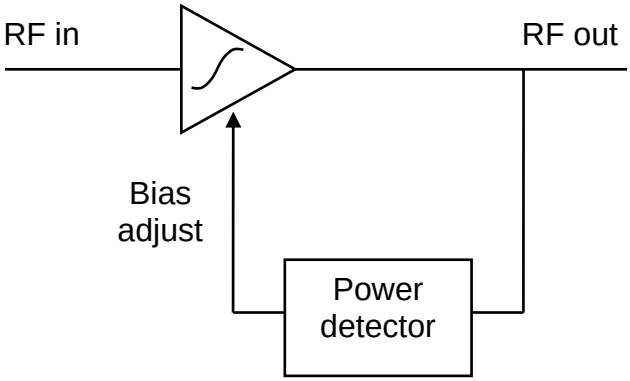


Figure 3.5: Block diagram of the bias adaptation technique

3.3 Linearization Techniques

As the name implies, linearization techniques attempt to improve the linearity of an amplifier, which is usually nonlinear. In some cases, a linearization technique is used with a linear amplifier in order to get a few more dB of linear output power, which can translate to a significant amount of peak efficiency improvement. The major advantage of using a nonlinear amplifier is its high efficiency. An effective linearization technique must consume much less DC power than the amplifier itself in order not to degrade the overall transmitter efficiency. Since linearization is not the main focus of this thesis, this will be discussed only briefly. Readers can refer to [12],[13],[14] for more details on linearization techniques.

3.3.3 Feedback

Feedback has been widely used to improve the linearity of low-frequency analog circuits. However, it is not as effective to use direct feedback at radio

frequencies since high gain at RF is harder to come by and stability could be a serious issue. By taking advantage of the fact that signals used in most wireless communication systems reside only in a narrow band, it is possible to down-convert the RF signal and close the feedback loop at a much lower frequency. This technique is referred to as indirect feedback.

The two most common indirect feedback techniques are Polar feedback and Cartesian feedback. Polar feedback separates the down-converted signal into amplitude and phase, whereas Cartesian feedback separates the down-converted signal into two orthogonal components (I and Q). The major disadvantage of Polar feedback is that it requires two feedback loops, amplitude loop and phase loop, which cannot be easily matched because the two paths are not identical. In Cartesian feedback, the two signal paths (I and Q) are identical and therefore can be easily matched. Besides, most modern communication systems already separate the signal into I and Q channels. Therefore, Cartesian feedback does not require extra effort to separate the amplitude and phase information as in Polar feedback. A block diagram of a Cartesian feedback transmitter is shown in Figure 3.6. Note that most blocks in a Cartesian feedback transmitter, with the exception of the down-conversion mixer, already exist in most typical transmitters.

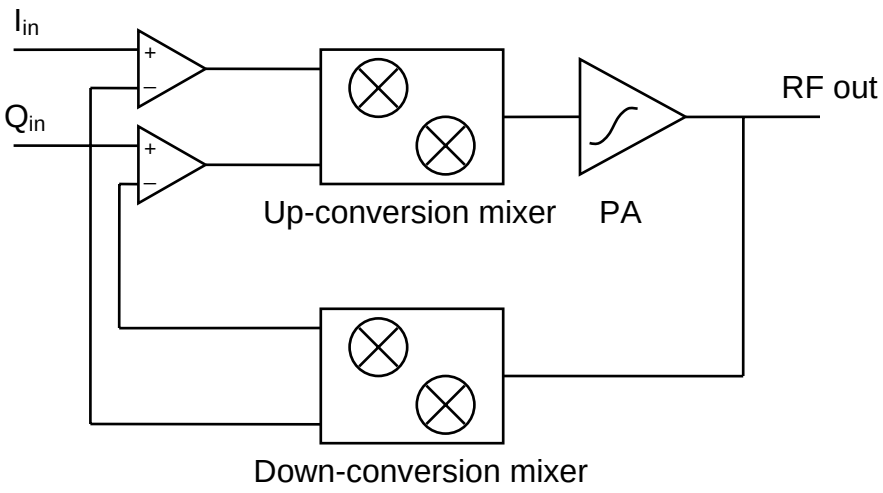


Figure 3.6: Block diagram of a Cartesian feedback transmitter

One important design aspect of a Cartesian feedback transmitter is that the down-conversion mixer must have high linearity, otherwise it will limit the overall linearity of the transmitter. A linear mixer can be power-hungry and is the power overhead to the Cartesian feedback system.

As in any feedback system, this feedback linearization scheme is subject to a stability problem. In order to compensate for the signal delay around the loop, a phase difference between the LO signals of the two mixers is deliberately introduced. The phase difference required to ensure stability is usually both signal-level- and load-dependent. Therefore, it must be designed with care to ensure stability at all output power levels and load values. A loop filter (not explicitly drawn in Figure 3.6) is required to set the loop bandwidth. Typically, it is very difficult to make the loop stable when the loop bandwidth is large. For this reason, Cartesian feedback is not suitable for applications with large signal bandwidth.

3.3.4 Predistortion

Predistortion is an open-loop technique and is not subject to stability problems. The predistortion technique predicts the nonlinearities of the amplifier and applies the inverse transfer function to either the baseband signal or the RF signal. By doing so, the overall transfer function of the transmitter can be made linear. In order to track changes in the amplifier characteristics due to temperature change or aging effect, the predistortion scheme can be made adaptive. The most common adaptive predistortion technique is used in the digital domain—hence the name digital adaptive predistortion. There are two major concerns in digital adaptive predistortion. First, due to the complex transfer characteristic of a nonlinear PA, the digital predistorter must be of high order and may have large power consumption. Second, since the predistorted signal has larger bandwidth than the original signal, the digital-to-analog converter used to convert the signal into the analog domain must have a much higher bandwidth. The bandwidth of the predistorted signal can be as large as four times the original bandwidth [15].

Due to its open loop nature, the predistortion technique achieves less precision than feedback techniques, and the predistorter must update its coefficients once in a while to track changes in the amplifier characteristic. This also translates to higher power consumption and lower overall efficiency of the entire transmitter chain. One significant advantage of this linearization technique is that there is not a limitation on the signal bandwidth that it can handle (except on DAC bandwidth). Besides, if predistortion is done digitally, the power

consumption and silicon area of the device will constantly benefit from technology scaling.

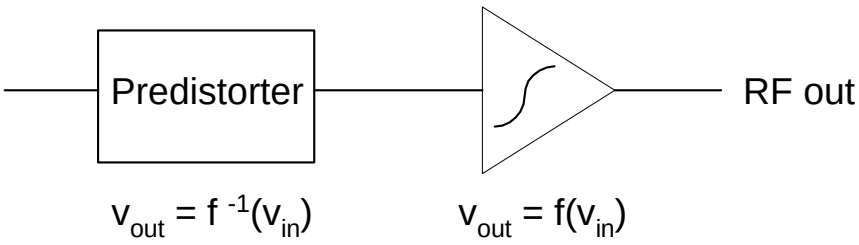


Figure 3.7: Predistortion technique

3.3.5 Envelope Elimination and Restoration (EER)

The EER technique, also known as the polar modulator, has been of much interest in recent years [16],[17]. Figure 3.8 shows a block diagram of an EER system. EER separates the input signal into two paths, amplitude path and phase path. The phase portion of the signal can be obtained by using a limiter. It is then amplified by a nonlinear PA. The amplitude portion of the signal can be obtained by using an envelope detector. This envelope information is then used to modulate the power supply of the PA via a DC-to-DC converter, thereby getting the amplitude information back at the output of the PA. This technique, like the power supply variation technique described in Section 3.2.2, uses a DC-DC converter. However, in EER the output voltage of the DC-DC converter must be able to change quickly enough to track the envelope of the signal. Since it is very difficult to design an efficient DC-DC converter with large bandwidth, this technique is generally limited to applications with relatively small bandwidth.

Another important characteristic that makes this technique unfit for applications with large bandwidth is that the phase delay of the two dissimilar signal paths cannot be easily matched.

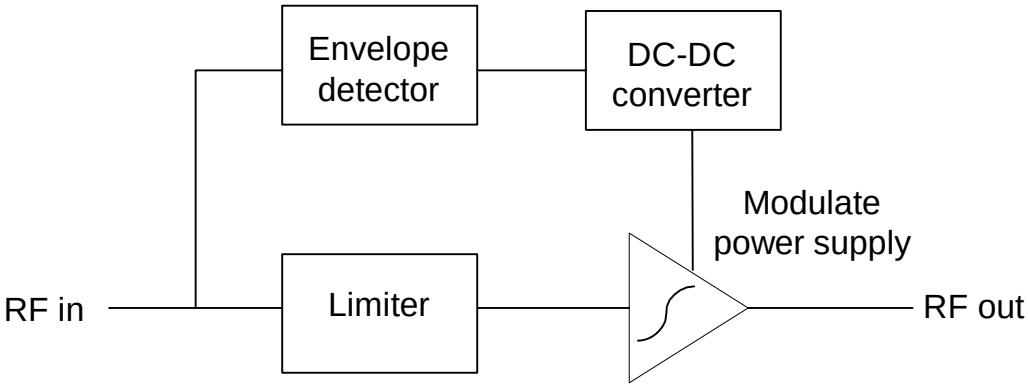


Figure 3.8: Block diagram of an EER transmitter

3.3.6 Chireix's Outphasing

The outphasing technique is also known as LINC (linear amplification by nonlinear components). It was invented in 1935 by Chireix [18]. The main idea of this technique is to decompose a non-constant envelope signal into two constant envelope signals.

$$x(t) = A(t)\cos(\omega_o t + \phi(t)) \tag{3.2}$$

$$= \cos(\cos^{-1} A(t))\cos(\omega_o t + \phi(t)) \tag{3.3}$$

$$= x_1(t) + x_2(t) \tag{3.4}$$

where

$$x_1(t) = \cos(\omega_o t + \phi(t) + \cos^{-1} A(t)) \tag{3.5}$$

$$x_2(t) = \cos(\omega_o t + \phi(t) - \cos^{-1} A(t)) \tag{3.6}$$

Note that in the above decomposition, A(t) is always assumed to be less than one; otherwise a normalizing factor has to be used. This decomposition can be depicted as in Figure 3.9. It can be seen that by phasing the two unit vectors $x_1(t)$ and $x_2(t)$ properly, the resultant vector can be of any phase with an amplitude between zero and two.

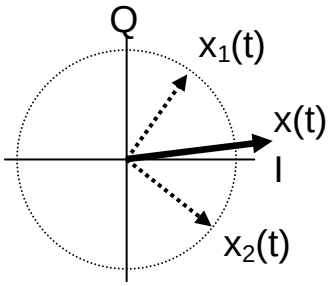


Figure 3.9: Outphasing decomposition

Since $x_1(t)$ and $x_2(t)$ both have constant envelope, two nonlinear amplifiers can be used. A block diagram is given in Figure 3.10. Even though this technique seems very attractive, there has yet to be found an efficient way of combining the output power of the two amplifiers. There is a major difference between this technique and the Doherty amplifier in regard to their power combining. In the Doherty amplifier, the phase difference between the main and auxiliary amplifiers is fixed at 90°, whereas in the outphasing technique the phase difference between the two nonlinear amplifiers is arbitrary. A microwave power combiner can be used, but it usually has high insertion loss and therefore significantly degrades the overall efficiency.

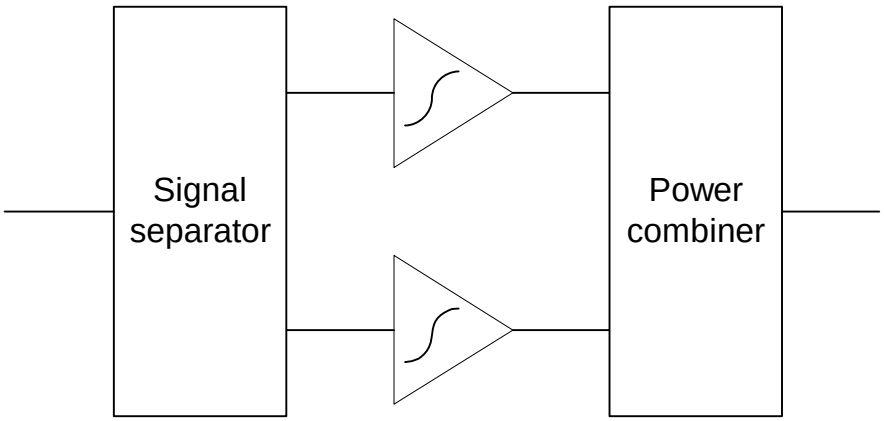


Figure 3.10: Block diagram of the outphasing technique

Chapter 4

The Doherty Amplifier

This chapter is dedicated to explaining the operation of the Doherty amplifier and its practical implementation issues. The linearity of the overall amplifier will be analyzed. The Doherty amplifier can also be used in nonlinear applications, and these will be comprehensively explained. In this technique, a passive impedance inverter (Z_{inv}) is used to efficiently combine the output power of the main and the auxiliary amplifiers. Instead of applying the quarter-wavelength transmission line that is conventionally used, a lumped element network can be used instead [10]. The foremost advantage of the lumped network is that it can be compactly integrated on-chip. However due to the low-Q on-chip inductor, the network loss becomes significant. The effect of inductor Q on the Doherty amplifier's operation and overall efficiency will be carefully analyzed. An algorithm that can be used to tune the Z_{inv} network to the desired operating frequency will be proposed. At the end of the chapter, the discussion will focus on how to extend the Doherty concept beyond two stages in order to

get high efficiency in an even wider output power range. Practical considerations in implementing a multi-stage design will be presented.

4.1 A Doherty Amplifier Block Diagram

A block diagram of a Doherty amplifier is shown in Figure 4.1. In this technique, a passive impedance inverter (Z_{inv}) is used to efficiently combine the output power of the two amplifiers and is placed at the output of the main amplifier. A Z_{inv} network inherently gives a 90° phase shift (or -90° depending on the topology used), so an additional 90° (or -90°) phase shift has to be added at the input of the auxiliary amplifier to match the delay of the two signal paths. A gain control block is required to control the amplitude of the input signal to the auxiliary amplifier in order to ensure proper operation of the Doherty amplifier. This gain control can also be done by adjusting the bias condition of the auxiliary amplifier itself.

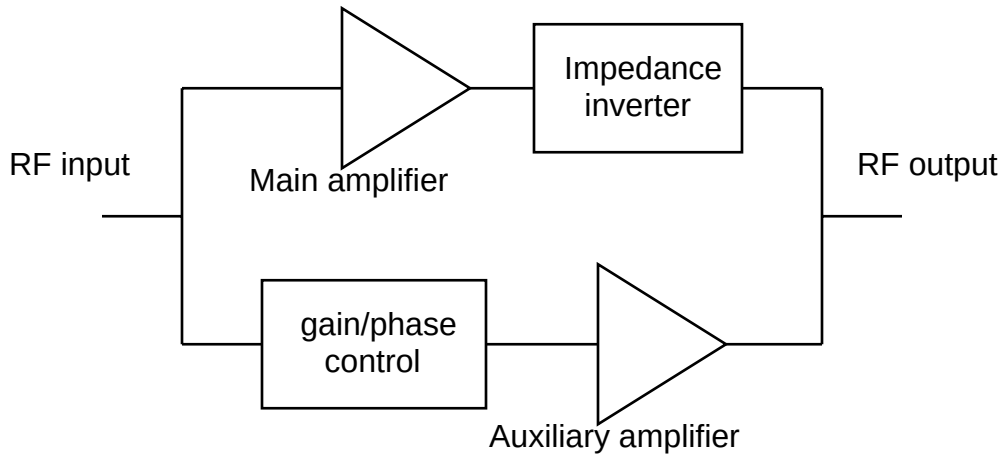


Figure 4.1: Block diagram of a Doherty amplifier

4.2 Passive Impedance Inverter (Z_{inv})

The passive impedance inverter used in the Doherty architecture is a symmetric network, which has the input impedance looking into one side of the network proportional to the reciprocal of the impedance on the other side of the network, scaled by a constant. In the original Doherty amplifier paper [10], a quarter-wavelength transmission line was used as the impedance inverter. However, at low GHz frequencies, the length of a quarter-wavelength transmission line is still on the order of a few centimeters and is not suitable for integration. An alternative is to use a lumped-element equivalent network as shown in Figure 4.2. Note that the networks in Figure 4.2b and 4.2d are the T-network equivalents of the pi networks in Figure 4.2a and 4.2c, respectively.

For all these networks, the impedance looking into one side of the network, Z_{in}, is

$$Z_{in} = \frac{X^2}{Z_{out}} \tag{4.1}$$

where Z_{out} is the impedance on the opposite side of the network and X is called the characteristic impedance of the network, similar to the characteristic impedance of a quarter-wavelength transmission line. At resonance, these networks and a quarter-wavelength transmission line are equivalent, as they all have exactly the same S-parameter matrix, which is

$$S = \begin{pmatrix} 0 & e^{\pm j90^\circ} \\ e^{\pm j90^\circ} & 0 \end{pmatrix} \tag{4.2}$$

Even though both distributed and lumped element networks look identical at the center frequency, they behave differently at harmonic frequencies and may affect the overall operation.

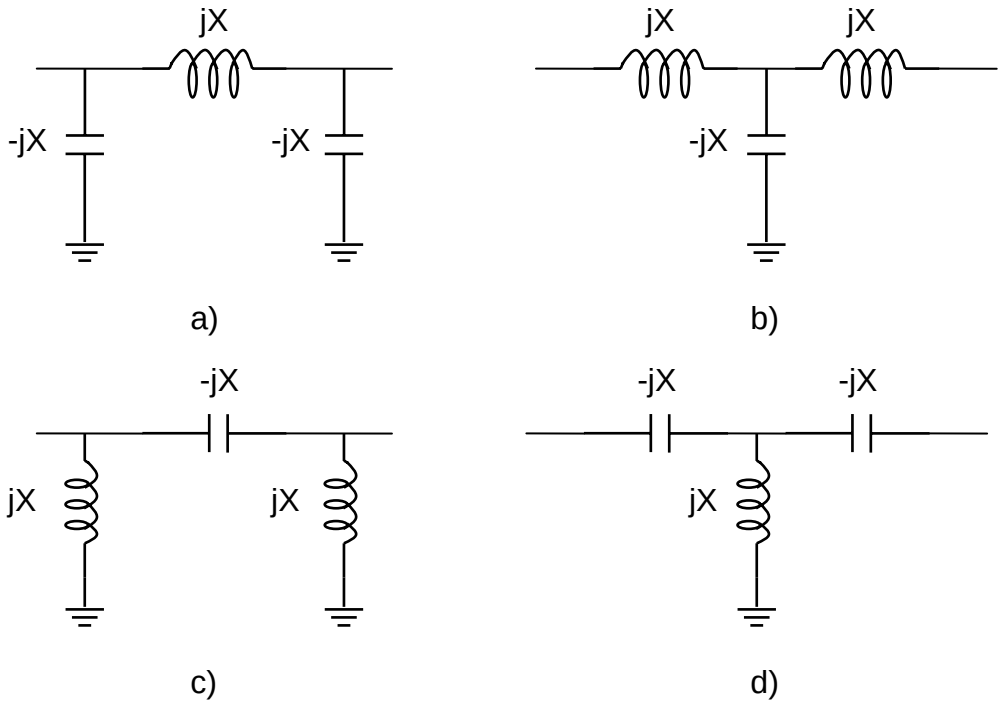


Figure 4.2: Lumped-element impedance inverters

4.3 Doherty Amplifier Operation

The operation of a Doherty amplifier can be understood by treating both amplifiers as voltage-controlled current sources. A simplified diagram is shown in Figure 4.3. In this diagram, it is assumed that the characteristic impedance of the impedance inverter is k times the load resistance (with k>1). During low output power operation, the auxiliary amplifier is off and the impedance seen by the main amplifier is

$$Z_m = \frac{(k R_L)^2}{R_L} = k^2 R_L \tag{4.3}$$

Since the impedance inverter network is assumed to be lossless here, by conservation of energy, it can be found that

$$I_m{}^2 (k^2 R_L) = I_o{}^2 (R_L) \quad \Rightarrow \quad I_o = k I_m \tag{4.4}$$

Using Equations 4.3 and 4.4, V_m and V_o can be found to be

$$V_m = I_m Z_m = (k^2 R_L) I_m \tag{4.5}$$

$$V_o = I_o R_L = (k R_L) I_m = \frac{V_m}{k} \tag{4.6}$$

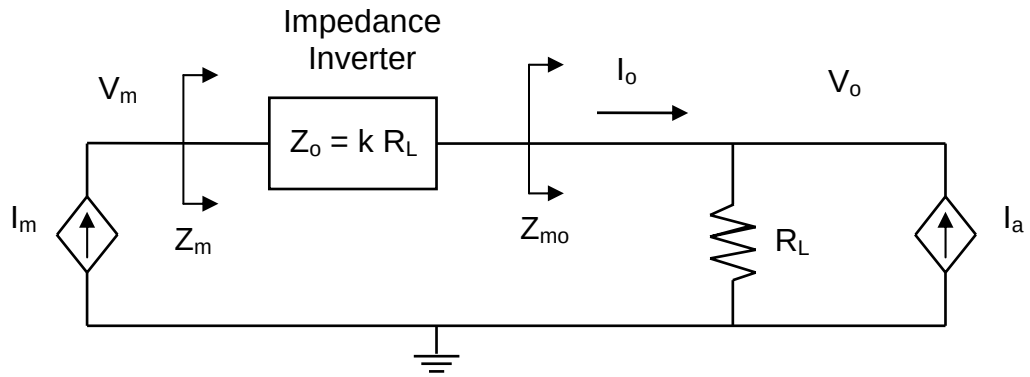


Figure 4.3: Simplified block diagram of a Doherty amplifier

Once the output of the main amplifier reaches its maximum swing, peak efficiency is obtained. At this point, the voltage swing at V_o is only 1/k times that of V_m. When more output power is needed, the auxiliary amplifier is also turned on. By adjusting the phase of I_a to be the same as I_o, the apparent impedance on the right side of the impedance inverter, Z_{mo}, becomes

$$Z_{mo} = \frac{V_o}{I_o} = R_L\left(1+\frac{I_a}{I_o}\right) \quad (4.7)$$

According to Equation 4.7, once the auxiliary amplifier is turned on, the impedance Z_{mo} increases (with I_a and I_o in phase). And by impedance inversion, the load resistance seen by the main amplifier, Z_m, decreases. Since Z_m is reduced, I_m can then be increased by the same proportion in order to keep V_m at the maximum swing. Once the auxiliary amplifier is on, by using the superposition theorem, Equation 4.4 can be re-written as

$$I_o = k\,I_m - I_a \quad (4.8)$$

The second term in Equation 4.8 arises from the fact that the impedance that the auxiliary amplifier sees is essentially a short circuit since the other side of the impedance inverter sees infinite output impedance from the main amplifier. Therefore, I_a is entirely absorbed into the Z_{inv} network rather than going to R_L. By applying this result to Equation 4.7,

$$Z_{mo} = R_L\left(1+\frac{I_a}{kI_m - I_a}\right) = R_L\left(\frac{k\,I_m}{k\,I_m - I_a}\right) \quad (4.9)$$

From this result, Z_m, V_m and V_o can be written as

$$Z_m = \frac{(k\,R_L)^2}{Z_{mo}} = \frac{(kI_m - I_a)(k)\,R_L}{I_m} \quad (4.10)$$

$$V_m = I_m\,Z_m = (kI_m - I_a)(k)\,R_L \quad (4.11)$$

$$V_o = (I_1 + I_o)\,R_L = (k\,R_L)\,I_m \quad (4.12)$$

An important observation drawn from the above equations is that the output voltage, V_o, depends only on the current from the main amplifier whether

or not the auxiliary amplifier is turned on, as long as the output of the main amplifier does not saturate. The main function of the auxiliary amplifier is to keep the main amplifier from exceeding its maximum output voltage swing, thus preventing it from saturating.

As the auxiliary amplifier is turned on, by proper controlling of I_m and I_a, the output power is increased while keeping the voltage at the output of the main amplifier at maximum swing. This condition can be achieved by taking the derivative of Equation 4.11 and setting it to zero.

$$\partial V_m = (k\,\partial I_m - \partial I_a)(k) R_L = 0$$

$$\Rightarrow \quad \frac{\partial I_a}{\partial I_m} = k \qquad (4.13)$$

Equation 4.13 indicates that once the auxiliary amplifier is on, its transconductance must be k times that of the main amplifier in order to keep V_m constant. Figure 4.4 shows voltage and current plots of the Doherty amplifier.

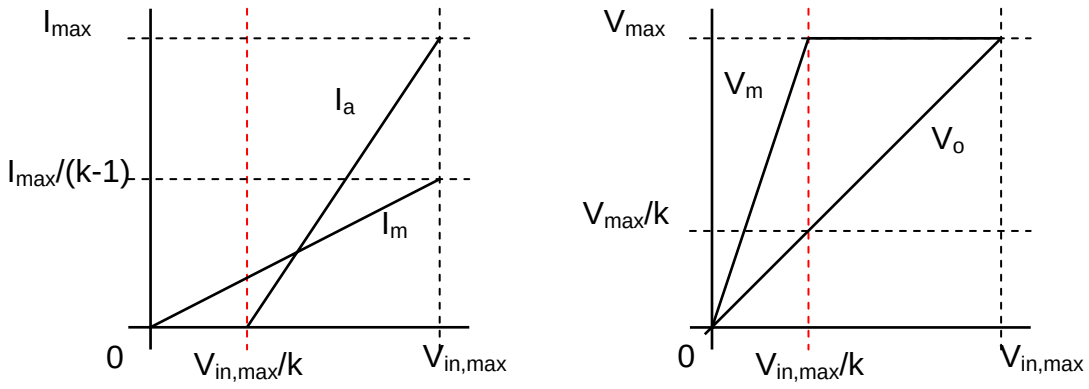


Figure 4.4: Voltage and current plots of the Doherty amplifier

At the breakpoint where the auxiliary amplifier is turned on, V_o is V_{max}/k, where V_{max} is the maximum voltage swing of both amplifiers. This implies that at the moment when the auxiliary is turned on, its efficiency is less than its peak value, causing the overall efficiency to drop. As both I_m and I_a increase, V_o also increases while V_m is constant. Once both V_o and V_m are at V_{max}, both amplifiers operate at their peak efficiency conditions, yielding the second peak in the overall efficiency plot. The overall efficiency of an ideal Doherty amplifier is shown in Figure 4.5.

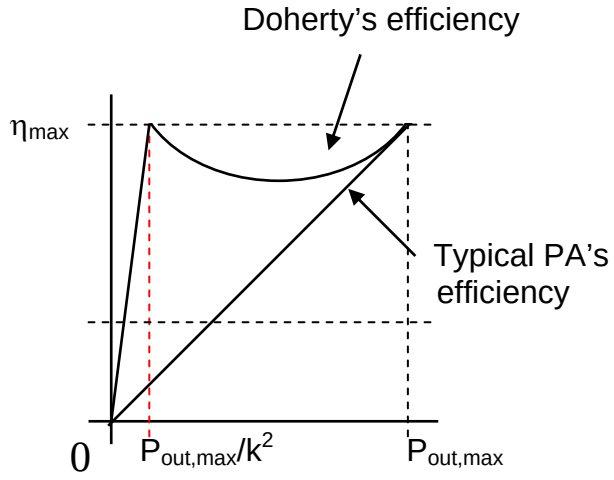


Figure 4.5: Overall efficiency plot of a Doherty amplifier

At the second efficiency peak, V_m is the same as V_o, Under this condition, Z_{mo} in Equation 4.9 and Z_m in Equation 4.10 have to be the same (conservation of energy on the left and right side of the impedance inverter). By using Equations 4.9 and 4.10, the following result is obtained:

$$I_a = (k-1)\, I_m \qquad (4.14)$$

The above equation shows the relative peak current value of the two amplifiers at peak power.

Another way to understand intuitively how a Doherty amplifier operates is to recognize that the impedance inverter gives a 90° phase shift and that the signals from the main and auxiliary amplifiers travel in the opposite direction through the impedance inverter network. By choosing the phase of I_a to be the same that of I_1, the voltage at V_m that is generated by I_a would be out of phase to that generated by I_m, thus preventing the main amplifier from exceeding its maximum voltage swing.

The output power of each amplifier can be found by multiplying the output current to the terminal voltage. Figure 4.6 shows the output power from the main and auxiliary amplifiers.

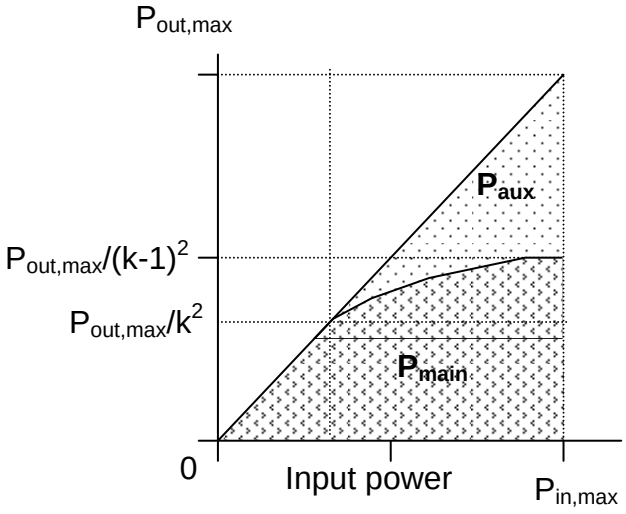


Figure 4.6: Output power from main and auxiliary amplifiers

In order to match the phase of the two amplifier paths, additional 90° phase shift (or -90°) must be inserted into the auxiliary amplifier path. An amplitude adjustment circuit is also needed to control the input of the auxiliary amplifier such that I_a follows the characteristic shown in Figure 4.4. As shown in Equation 4.12, the output of the Doherty amplifier depends only on the output current of the main amplifier. Therefore, the only design criterion for the main amplifier is to make it linear enough to meet the linearity specifications. As for the auxiliary amplifier, since it does not affect the linearity, the design strategy there is to make it as efficient as possible while still keeping the main amplifier from saturating. By designing the auxiliary amplifier to be more efficient than the main amplifier, the second peak of the overall efficiency, which is the weighted average of the efficiencies of the two amplifiers, can potentially be higher than the first peak. Typically, a class A, AB, or B PA is used to realize the main amplifier, depending on the level of linearity required. A class C is traditionally used to implement the auxiliary amplifier. A switching amplifier cannot be used because its output current only weakly depends on its input voltage, thus making it impossible to get the output current characteristic shown in the Figure.

When using a class C PA as the auxiliary amplifier, there are a few issues to be considered. In order for a class C amplifier to generate the same level of output current as a class A, AB, or B amplifier, the device size must be much larger and might not be feasible. This is especially true when the peak current from the auxiliary amplifier has to be many times larger than that from the main amplifier (when k is much larger than 1). And since a class C amplifier is

inherently nonlinear, it is not possible to get the linear input-output response shown in Figure 4.4 without adjusting its bias point or its input drive. Figure 4.7 shows the typical voltage transfer characteristic and efficiency of a class C amplifier. To use this class C amplifier as the auxiliary amplifier, the active device must be sized such that its output current curve is above the ideal curve across the whole range of all output voltage in order to keep the main amplifier from saturating. The resulting current, voltage, and efficiency plots are summarized in Figure 4.8. In this example, it is assumed that both amplifiers have linear instantaneous I-V characteristic. Note that the efficiency at the second peak is not optimal since V_o is not at the maximum swing.

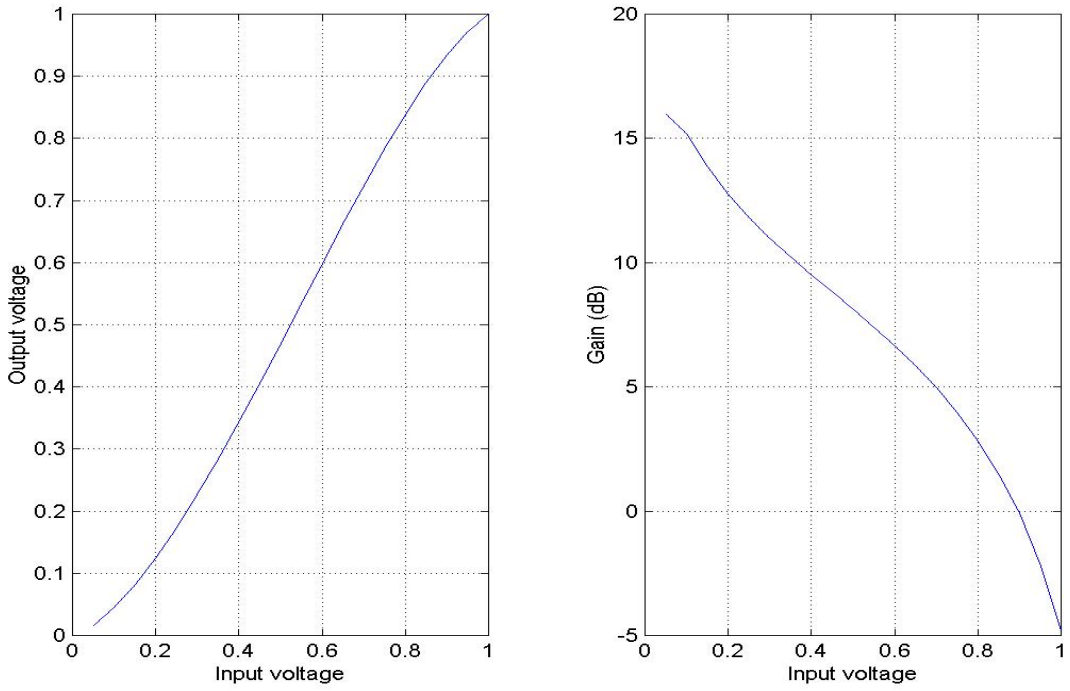


Figure 4.7: Typical voltage transfer characteristic of a class C amplifier

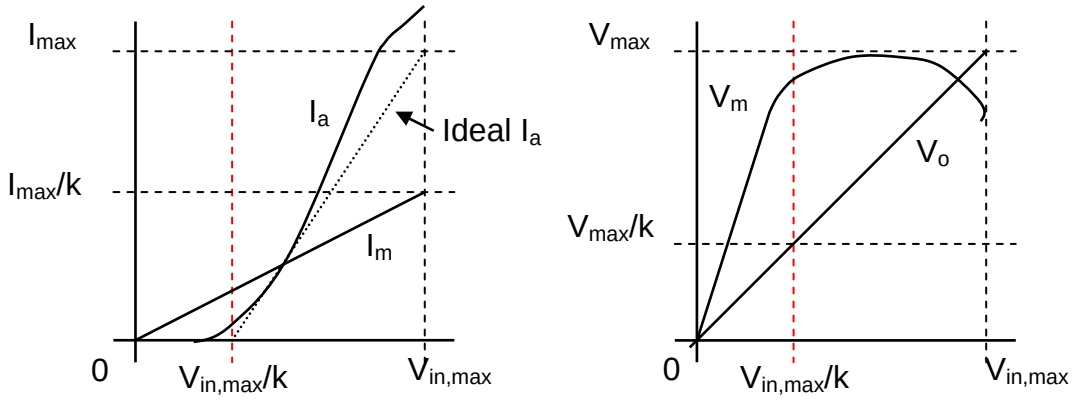


Figure 4.8: Doherty amplifier characteristic with a class C auxiliary amplifier

To mitigate this problem, a gain adjustment circuit can be applied. This can be done by means of bias adjustment and/or changing the gain of the driving stage. It might involve more complicated circuitry, but might also prove to be worthwhile.

To turn off the auxiliary amplifier below the break point, one obvious solution would be to use a class C amplifier and set its bias point such that the amplifier is turned off below the breakpoint without aid from the amplitude adjustment circuit. This technique is sometimes referred to as a self-biased Doherty amplifier. However, in order to do this, the class C amplifier must be biased in the deep class C region, which would require an even larger active device in order to realize the desired output current level. An alternative approach is to use the information from the power control signal to determine whether the auxiliary amplifier is needed, and turn it on and off as necessary.

Another important caveat about the Doherty amplifier is that the plots in Figure 4.4 are for the fundamental component of the signal. The I-V

characteristic in Figure 4.4 has been mistaken by many as being the instantaneous characteristic, requiring that a class C amplifier be used as the auxiliary amplifier. This is inaccurate since the Z_{inv} network does not have infinite bandwidth. Therefore the operation is valid only in a narrow frequency band. In fact, both amplifiers can very well be class B, or even class A. With proper gain adjustment, all the explanations and equations above still hold.

4.4 The Linear Doherty Amplifier

Figure 4.9 is a simplified diagram of the Doherty amplifier. The main and auxiliary amplifiers are represented by two current sources, I_m and $I_{aux,}$ respectively. The Z_{inv} network is assumed to have a +90° phase shift here (all the analysis that follows also holds for a -90° phase shift Z_{inv} network). The linearity of the Doherty amplifier can be best understood by using the superposition theorem. The first scenario is when the main amplifier is on and the auxiliary amplifier is off. This yields:

$$V_{m,main} = I_m (k^2 R_L) \tag{4.15}$$

$$V_{o,main} = j\frac{V_{m,main}}{k} \tag{4.16}$$

Note that the subscript *main* indicates the contribution from the main amplifier when the superposition theorem is used (and the same goes for subscript *aux* which will be used later).

The second scenario is when the main amplifier is off and the auxiliary amplifier is on. If the main amplifier is assumed to have infinite output impedance, it looks like an open circuit once it is turned off. By impedance inversion, the

impedance looking into the right side of this network is a short circuit, under the assumption that the Z_{inv} network has infinite Q. Therefore, the current from the auxiliary amplifier is entirely absorbed by the Z_{inv} network and does not produce any voltage at the output node. However, it does create a voltage swing at V_m given by

$$V_{m,aux} = -I_{aux}(k R_L) \tag{4.17}$$

The minus sign in the above equation arises from the fact that the auxiliary amplifier current has +90° phase compared to the main amplifier current. And once the signal travels through the Z_{inv} network, it experiences another +90° phase shift and therefore produces a negative voltage at V_m. This, in fact, is another way to understand the operation of the Doherty amplifier. Due to signal cancellation at V_m, the main amplifier can output more current while keeping its terminal voltage constant. With regard to the linearity of the overall amplifier, since the auxiliary amplifier current does not affect the output voltage, this implies that any in-band nonlinearities from the auxiliary amplifier do not affect the overall linearity.

In the spectrum plots in Figure 4.9, solid lines represent the contribution from the main amplifier and dotted lines represent that of the auxiliary amplifier. In this Figure, it is assumed that the desired signal is at frequency f_o and the component at f_1 from the auxiliary amplifier is an odd-order intermodulation product created by two closely spaced tones around f_o (only one of them is shown in Figure 4.9 for simplicity). Both f_o and f_1 are assumed to be within the bandwidth of the impedance inverter. Even though the component at f_1 does not

affect V_o, it does have some effect on V_m. If the main amplifier is an ideal transconductor, this additional voltage generated at V_m will not affect I_m and the overall linearity of the Doherty amplifier. Another way to visualize this is to convert I_{main} and Z_{inv} into a Thevenin equivalent as shown in Figure 4.10. It can be clearly seen that I_{aux} does not have any effect on V_o, since V_o is controlled by a voltage source that depends only I_{main}.

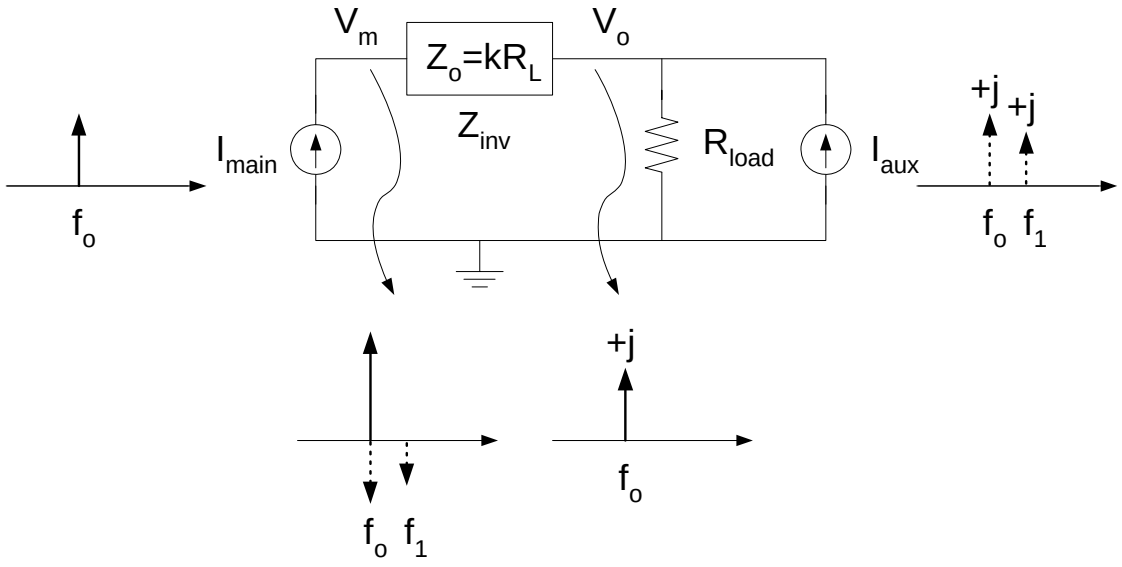


Figure 4.9: Simplified diagram of a Doherty amplifier

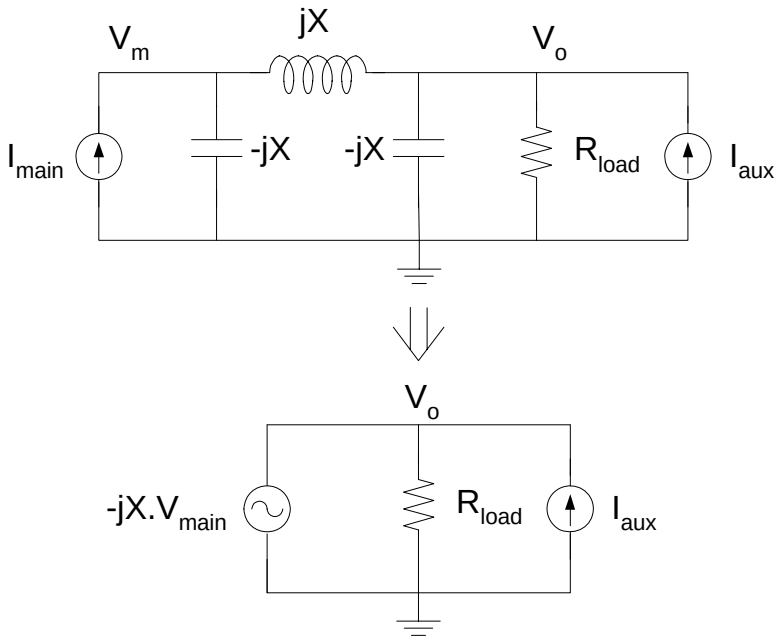


Figure 4.10: A simplified Doherty amplifier diagram with Thevenin equivalence

From the discussion above, it can be concluded that ideally the linearity of a Doherty amplifier depends only on the main amplifier, whereas the overall efficiency is a weighted average of both amplifiers. Therefore, if there is a linearity target that a Doherty amplifier has to meet, the main amplifier has to be designed accordingly and the auxiliary amplifier can be less linear to improve the overall efficiency. This implies that, even for constant envelope signals, the second peak of the overall efficiency plot has to be higher than the first one unless both amplifiers are made equally efficient.

However, in reality, the Q of the Z_{inv} network and the output impedance of an amplifier are not infinite. Therefore, under the superposition theorem, the auxiliary amplifier actually sees a small but finite impedance looking into the Z_{inv} network. As a result, the in-band nonlinear components (at frequency f_1 as in the

previous example) can still appear across the load. This dictates how nonlinear an auxiliary amplifier can be. Even with finite Q of the Z_{inv} network, one should still observe the second efficiency peak to be higher than the first peak.

4.5 Effect of a Lossy Z_{inv} Network

For all the discussions above, it was assumed that the Z_{inv} network is lossless. This is not usually the case in practice, especially when on-chip components are used. This section discusses the effect of a lossy Z_{inv} network on the operation and overall efficiency of a Doherty amplifier and the details of its power efficiency. In this section, the topology in Figure 4.2a is used for the Z_{inv} network. It is also assumed that the series component (an inductor in this case) has finite Q, whereas the other two shunt components have infinite Q. If the shunt components have finite Q, their parallel-represented parasitic resistance can be lumped together with the load resistance or the parasitic resistance from the inductors at the output of the amplifiers.

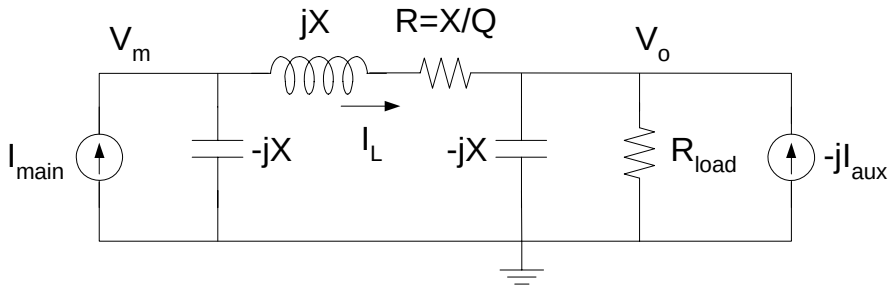


Figure 4.11: Simplified diagram of a Doherty amplifier with a lossy Z_{inv} network

From the diagram in Figure 4.1, the phase shift across the lossy Z_{inv} network was calculated to be

$$\arg\left(\frac{V_O}{V_m}\right) = \arg\left(\frac{-jX \,//\, R_{load}}{R + jX + (-jX \,//\, R_{load})}\right)$$

$$= \arg\left(\frac{-jX\, R_{load}}{R_{load}\,(1+\frac{j}{Q}) + R}\right)$$

$$= -90° + \arg\left(\frac{R_{load}}{R_{load}\,(1+\frac{j}{Q}) + R}\right) \qquad (4.18)$$

With finite Q, the phase shift is no longer 90°. And since the effective R_{load} changes as the auxiliary amplifier is turned on, the phase shift across the Z_{inv} network is signal-level dependent. After a lengthy mathematical manipulation, V_m and V_o in Figure 4.11 were found to be

$$V_m = \frac{R_{load}}{R_{load}\,(1+\frac{j}{Q}) + R}\left[\frac{X^2}{R_{load}}\left(1+\frac{R_{load}}{X\,Q} - j\frac{1}{Q}\right) I_{main} - X\, I_{aux}\right] \qquad (4.19)$$

$$V_O = \frac{R_{load}}{R_{load}\,(1+\frac{j}{Q}) + R}\left(-jX\right)\left(I_{main} + \frac{I_{aux}}{Q}\right) \qquad (4.20)$$

For R=0 (Q=∞), Equations 4.19 and 4.20 reduce to Equations 4.11 and 4.12, respectively, as expected. Equations 4.19 and 4.20 imply that with R>0, both amplifiers see complex impedance, unlike the lossless Z_{inv} case. This degrades the efficiency of both amplifiers. The term within the brackets in Equation 4.19 indicates that voltage subtraction at the output of the main amplifier is not in

phase anymore. As a result, I_{aux} has to be increased in order to keep the main amplifier from saturating, or else the linearity of the main amplifier could be greatly compromised.

With a lossless Z_{inv} network, the ratio of I_{main} to I_{aux} at full power is (k-1) where $k = X/R_{load}$. This is no longer accurate with a lossy Z_{inv} network. At full power, the magnitude of both V_m and V_o should be at the maximum swing for best efficiency. The ratio of I_{main} to I_{aux} at full power can be found by equating the magnitudes of the right hand sides of both Equations 4.19 and 4.20.

$$\left\| (-jX)\left(I_{main} + \frac{I_{aux}}{Q} \right) \right\| = \left\| \frac{X^2}{R_{load}} \left(1 + \frac{R_{load}}{X\,Q} - j\frac{1}{Q} \right) I_{main} - X\, I_{aux} \right\| \qquad (4.21)$$

By setting $X=kR_{load}$ and ignoring the second term on the right hand side (assuming $kQ << 1$),

$$k^2 R_{load}{}^2 \left(I_{main} + \frac{I_{aux}}{Q} \right)^2 = \left(k^2 R_{load}\, I_{main} - kR_{load} I_{aux} \right)^2 + \left(k^2 R_{load} \frac{I_{main}}{Q} \right)^2 \qquad (4.22)$$

From the above equation, the ratio I_{aux}/I_{main} can be obtained by solving a quadratic equation. Both obtained solutions give $|V_m|=|V_o|$ but the larger solution reflects the case in which the auxiliary amplifier current is large enough to cause 180° phase inversion of V_m. When this happens, the main amplifier no longer outputs power, but instead dissipates power that must then be supplied by the auxiliary amplifier. Therefore, this solution is not power efficient. The other solution was found to be

$$\frac{I_{aux}}{I_{main}} = \frac{\left(k + \frac{1}{Q} \right) - \sqrt{1 + \frac{2k}{Q}}}{\left(1 - \frac{1}{Q^2} \right)} \qquad (4.23)$$

with $Q=\infty$, $I_{aux}/I_{main}=k-1$ as expected. But with finite Q, the amplifier current ratio decreases. For example, with k=3 and Q=8, I_{aux}/I_{main} is 1.83 rather than 2 in an ideal case.

4.6 Tuning of a Z_{inv} Network

To allow a Doherty amplifier to work properly, the magnitude of impedance of all three reactive components of the Z_{inv} network must be matched. Switched capacitor arrays can be used in place of the two shunt capacitors and act as the tuning elements. A tuning algorithm will now be presented; the task can possibly be done automatically.

Tuning can be done by turning on only the auxiliary amplifier, as shown in Figure 4.12. In this Figure, the bondwire inductance and the output capacitance of the main amplifier are lumped into X_1 of the Z_{inv} network (variation in bondwire inductance can also be tuned out by this tuning scheme). All the shunt elements at the output of the auxiliary amplifier are lumped together as Z_{load}. If X_2/Q_L is small, a relatively simple tuning scheme can be used. In this scheme, only the auxiliary amplifier is turned on and the value of X_1 is swept across its tuning range. The desired capacitor code is attained when the amplitude of the voltage swing at V_o is the smallest, which is when X_1 and X_2 resonate (an amplitude detector at V_o is needed). However, since the output capacitance of the main amplifier changes with the output swing due to nonlinear junction capacitance, a more accurate tuning scheme can be applied. With another amplitude detector at V_M, the input amplitude of the auxiliary amplifier can be increased until the signal

swing at V_M reaches maximum value. Doing this makes the output capacitance of the main amplifier approximately the same as when it is supplying full output power.

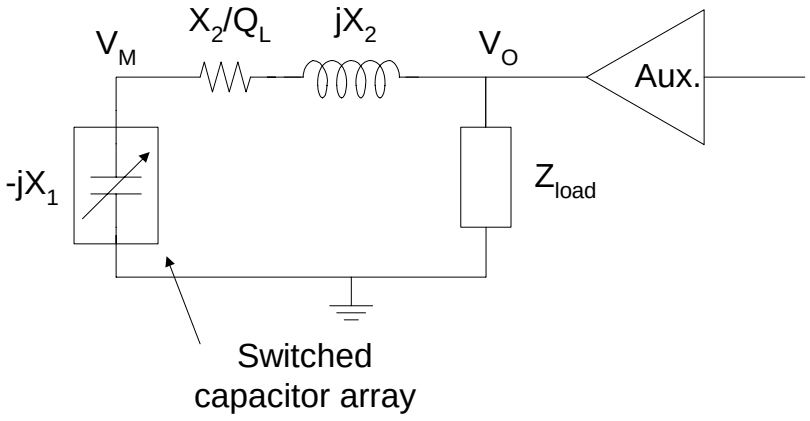


Figure 4.12: Tuning of the Z_{inv} network

However, if Q_L is not high enough or the imaginary part of Z_{load} is small, the algorithm stated above will give an inaccurate solution. This can be understood by looking at the overall admittance of the passive components at the output of the auxiliary amplifier, Y_{aux}. Figure 4.13 shows Y_{aux} plotted on a Y-plane. Note that the admittance of the leftmost series branch follows a circular trajectory as shown by the dotted line. From Figure 4.13 it can be seen that if R and/or R_{load} is much smaller compared to X_{load}, the maximum Y_{aux} and the desired Y_{aux} are approximately the same. But if that is not the case, a second capacitor array can be used to adjust X_{load}. For each value of X_{load}, a maximum Y_{aux} can be found, and the desired Y_{aux} is the smallest maximum Y_{aux}. This condition can also be written as

$$Desired\ Y_{aux} = \min_{Xload}\left(\max_{X_1}\left(Y_{aux}\right)\right) \tag{4.24}$$

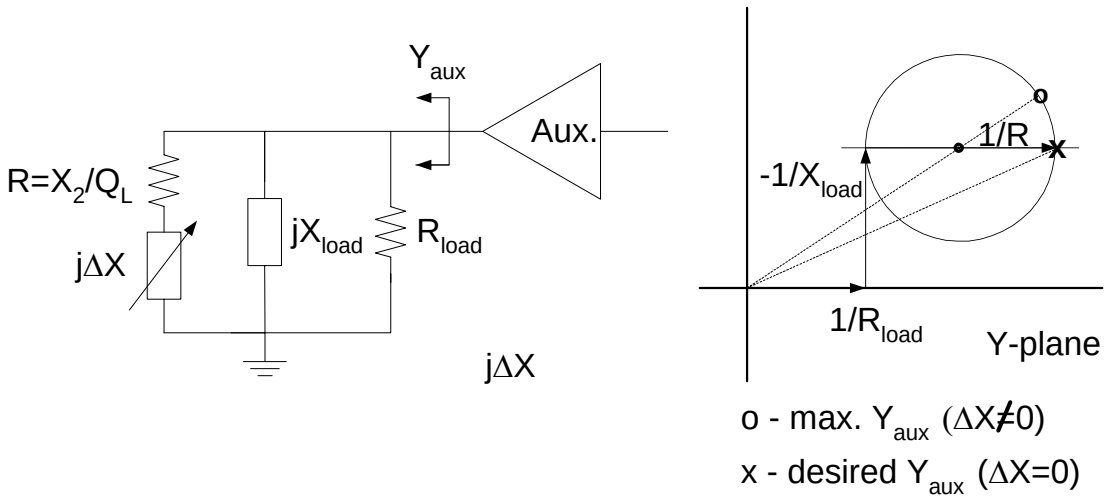


Figure 4.13: Y_{aux} plot on a Y-plane

Once the correct value of X_1 is found, the same capacitor code can be used with the second shunt capacitor array if both amplifiers, including their output bondwire inductance, are identical. If they are not identical, a much more complicated tuning scheme will be needed. An example might be detection of the phase difference between V_M and V_o when only the main amplifier is on. Therefore, in order to simplify the tuning algorithm, both amplifiers should be designed to be identical. Besides, by having both amplifiers identical, the phases of the two signal paths can be better matched.

4.7 The Nonlinear Doherty Amplifier

Thus far, the discussions of the Doherty amplifier are based on the assumption that both amplifiers have high output impedance. Even though this is

suitable for a linear amplifier design, nonlinear ones such as class C amplifiers often does not have that property. Nonetheless, the Doherty amplifier technique can still be effectively applied to nonlinear cases. This section focuses mainly on using class C amplifiers in a Doherty configuration. Unlike other switching amplifier classes, the output of a class C amplifier can still be controlled by the level of the input signal. Therefore, a DC-DC converter is not needed to do power control in this case.

A CMOS class C amplifier is usually a transconductance amplifier at low output swing. But as the output swing gets larger, the output resistance of the active device in use drops and the output of the amplifier starts to look more like a voltage source. Figure 4.14 shows a Doherty amplifier with both sub-amplifiers biased as class C. In this figure, it is assumed that the main amplifier has already saturated and the auxiliary amplifier just turned on. In this circumstance, it is valid to assume that the auxiliary amplifier output impedance is still high.

The lower part of Figure 4.14 is obtained by converting V_{main} and the Z_{inv} network into a Norton equivalent. Instead of having V_o determined solely by the main amplifier output current as before, I_1 is now constant regardless of I_{aux} and V_o. Unlike the linear case, once I_{aux} is turned on, the input signal to the main amplifier does not have to change as long as it can keep the amplifier saturated all the time. The efficiency of the main amplifier is thus kept at the maximum value all the time since its output swing saturates. To obtain the second overall efficiency peak, I_{aux} has to be increased until V_o reaches the maximum swing. In this case, the linearity of both amplifiers affects the overall linearity. At full power,

both amplifiers saturate and have low output impedance, but the operation of the Doherty amplifier remains valid.

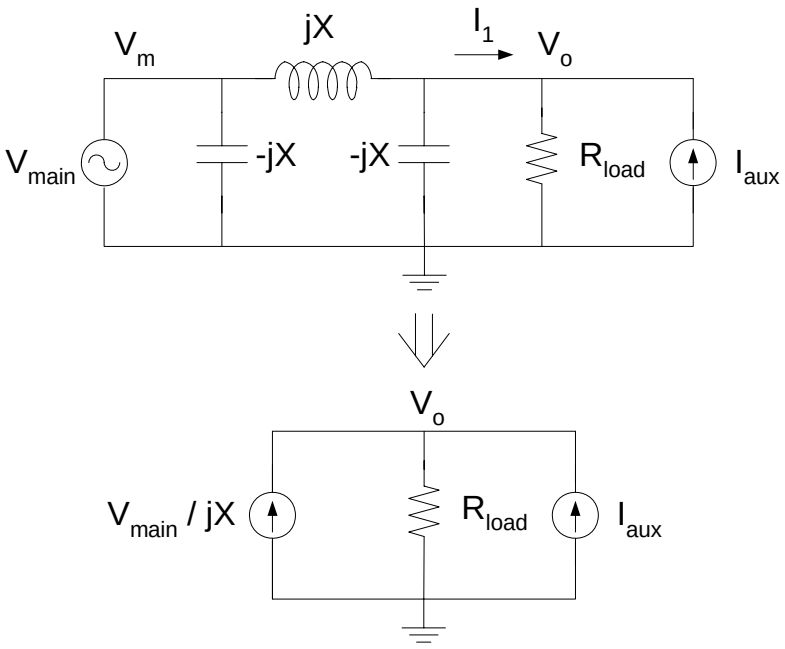


Figure 4.14: Simplified diagram of a Doherty amplifier using a Norton equivalent

As for switching amplifiers, their output swing is mainly determined by the power supply voltage. Power control for this type of amplifier is done by means of power supply variation (see Section 3.2.2). Therefore, it is not necessary to use the Doherty technique as another means of power control, even though it is possible in principle. For a fixed power supply, two efficiency peaks can be obtained at two power levels. In order to get other power levels, a DC-DC converter would be needed.

4.8 Effect of Load Variation

Under nominal conditions, the impedance of an antenna is 50Ω. However, as the environment around the antenna changes, its characteristic impedance can differ from 50Ω. With an output matching network designed for 50Ω antenna impedance, the load impedance presented to the amplifier (Z_{load}) can change significantly. The discussion in this section is based on the assumption that the Z_{inv} network is lossless, since this allows the effect of load variation to be intuitively understood.

The effect of load variation on the operation of a Doherty amplifier can be understood by looking at the admittance $Y_{load}=1/R_{load}+j/X_{load}$ rather than Z_{load}. First let us assume that X_{load} is much larger than R_{load}. If the value of R_{load} is different from its nominal value, the ratio of the characteristic impedance of the Z_{inv} network to R_{load} changes. This causes the breakpoint at which the auxiliary amplifier is supposed to turn on to change. The two peaks in the overall efficiency are still preserved but the first peak occurs at a different output power level. This also implies that the maximum current levels from both amplifiers at full power also change. The best way to detect this change in R_{load} is to use two amplitude detectors, one for each amplifier output, and let the auxiliary amplifier turn on once the output swing of the main amplifier reaches the maximum value. However, once the auxiliary amplifier is turned on, its gain has to be properly adjusted to prevent the main amplifier from saturating (in the linear case). In the nonlinear case, since the main amplifier is already saturated, the current from the

auxiliary amplifier has to be adjusted only to achieve the desired output power level.

If X_{load} is small but not negligible, its effect can be understood by looking at Equations 4.19 and 4.20, with Q=∞, R=0 and R_L being replaced by Z_{load}. This yields

$$V_M = \frac{X^2}{Z_{load}} I_{main} - X I_{aux} \quad (4.25)$$

$$V_o = -jX I_{main} \quad (4.26)$$

From the equations above it can be seen that the output voltage is not affected by complex Z_{load}. And since I_{main} is proportional to I_{aux}, the impedance that the auxiliary amplifier sees is still purely real. But the impedance that the main amplifier sees is now complex, and this degrades the main amplifier and the overall efficiency. With the presence of X_{load}, $|Z_{load}|$ becomes smaller and hence larger impedance is seen by the main amplifier. This implies that the current coming from the auxiliary amplifier has to be larger in order to prevent the main amplifier from exceeding its maximum output swing. It is better to use simulation to find out how much more the current from the auxiliary amplifier has to increase, since the subtraction at the output of the main amplifier is not in phase and therefore is not easy to analyze.

In the case where load variation is severe, an isolator might be needed at the PA output to guarantee proper operation of the Doherty amplifier.

4.9 The Multi-stage Doherty Amplifier

The concept of the Doherty amplifier can potentially be extended beyond two stages. With more stages, the efficiency at large back-off power can be further improved. This is particularly important in a system that has typical output power much lower than the peak.

From the discussions in the previous sections, it can be concluded that it does not matter whether the main amplifier has current source or voltage source output, because a Z_{inv} network and an additional amplifier (namely, an auxiliary amplifier, in the two-stage case) can be placed at the main amplifier output to get another efficiency peak. Therefore, in order to design an n-stage Doherty amplifier, one can start by designing a two-stage amplifier such that the first efficiency peak is at the lowest output power level and the second peak is at the second-lowest. This two-stage amplifier can be thought of as a new main amplifier. Then another set of Z_{inv} network and an auxiliary amplifier can be added as the third stage, which gives a third efficiency peak at a higher output power. With this higher output power level, the characteristic impedance of the second Z_{inv} network is lower than that of the first one. By recursively adding more stages, an n-stage Doherty amplifier can be achieved. Separations between two efficiency peaks (in term of output power) do not have to be uniform, depending on how the characteristic impedances of the Z_{inv} networks are designed. As an example, a four-stage Doherty amplifier is shown in Figure 4.15. These four stages are sized such that the two adjacent efficiency peaks are 6dB apart.

Even though it seems that having more stages is advantageous, its benefit could be outweighed by the fact that in practice a Z_{inv} network is lossy. Therefore, the efficiency peaks get progressively smaller as the output power decreases. Besides, the highest efficiency peak also degrades with an increasing number of stages. Typically, the effort to improve the PA efficiency is made under the assumption that it is the most power-hungry block in a transceiver. But at large output power back-off, this statement is not true any more. Therefore, trying to improve the efficiency of a PA at low output power level may improve the transmitter efficiency only marginally.

Chapter 5

Linear Amplifier Design

Wireless communications in today's world demands high data rates for various emerging applications. As bandwidth becomes increasingly scarce, the utilization of modulation schemes with high spectral efficiency is unavoidable. However, these spectral efficient modulation schemes carry information in both amplitude and phase. In order to preserve the integrity of the signal, a linear transceiver is required. The discussion of linear receivers is omitted in this thesis. Readers can refer to [19],[20] for more details. On the transmit side, the component that is the most difficult to design with high linearity is the power amplifier. One reason for this is that PAs usually have to deal with large signal swings rather than small signals as in other circuit blocks. Another reason is that PAs also have to achieve good efficiency, which usually is in conflict with achieving high linearity. Most other types of circuits can easily trade power efficiency for linearity without significantly affecting the overall transmitter efficiency.

The discussion in the chapter focuses mainly on a watt-level PA design. In deep submicron CMOS technology, the oxide breakdown voltage of a transistor is usually very small. A thick oxide transistor is frequently utilized as a cascode transistor to allow the use of higher power supply voltage, thus yielding higher output swing. A thick oxide transistor is commonly available in most modern CMOS processes for I/O purposes and usually has an oxide breakdown voltage in excess of 3V. To get high output power, an output matching network is needed to transform the impedance downward, as discussed previously. Since the output stage has to drive such a low impedance level, the transistor in the output stage must be large enough to handle the peak current. A transistor width of more than 10mm is not uncommon. With large transistor width, the input capacitance of the output stage is usually much too large to be driven by the preceding circuit. This requires a driving stage (or two in some cases) in order to reduce the overall input capacitance, and thus the input power, to a level that can be supplied by the preceding circuit block. Driver and predriver are typically designed to be linear even in a nonlinear PA since their power consumption is much smaller than that of the output stage, thus allowing the entire nonlinearity budget to be allocated to the output stage where the linearity-efficiency tradeoff is much more severe. Driver stage design will be covered briefly toward the end of this chapter. A fully differential design is assumed throughout the chapter.

5.1 Linear Output Stage Design

The output stage is the most critical part of PA design. The main function of the output stage is to provide high output power with a substantial amount of power gain while achieving good power efficiency. Typically, an amplifier stage has no more than 10-20dB of power gain. High power gain can potentially cause instability and is generally avoided. A fully differential topology is often preferred because it doubles the output swing and the load resistance (from $P=V^2/R$), and halves the current level. Higher load resistance relaxes the transformation ratio of the output matching stage and usually translates to lower insertion loss. All even harmonics are cancelled when the output is taken differentially. Besides, differential topology helps reduce the ground bounce since the common source node acts as a virtual ground to the fundamental component and all of the odd harmonics.

Conventional wisdom in linear PA design is to use a class A amplifier to avoid signal clipping and saturation of the active device. The implementation relies on the active device's being a linear transconductor and backing off from the maximum output swing until the linearity requirements are met. In short-channel CMOS, transistors do not behave as a good transconductor as in long-channel CMOS. The device characteristic deviates considerably from the square law and the output resistance is much lower. As a result, a class A amplifier has to back off from the maximum output voltage swing considerably in order to meet its linearity requirements. This often yields very poor efficiency—less than 5% in some cases where modulated signal with high peak-to-average ratio are used.

Even though a long-channel CMOS transistor behaves close to a linear transconductor, it has other shortcomings that make it difficult to supply high output power.

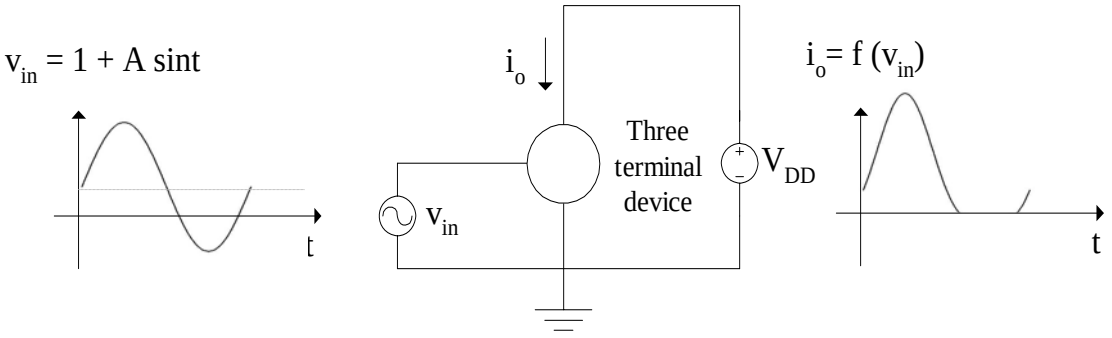
An alternative to a brute-force class A linear amplifier is to allow certain nonlinearities in the amplifier and balance them out to obtain a linear overall characteristic. This can be done by exploring a mechanism that causes gain expansion and attempting to cancel it out by another mechanism that causes gain compression. As a starting point, the class AB amplifier is examined. It is an attractive candidate since it still preserves a certain degree of linearity and its efficiency is much better than that of a class A. The class AB amplifier allows clipping of the input signal (the active device to turns off during some part of the cycle) and has a conduction angle between 180° and 360°. For the purpose of this discussion, let us assume that the active device in use has the I-V characteristic given by:

$$i_o = \begin{cases} f(v_{in}) & for\ v_{in} > 0 \\ 0 & for\ v_{in} < 0 \end{cases} \tag{5.1}$$

The input signal is assumed to be

$$v_{in} = 1 + A\sin t \tag{5.2}$$

Clipping occurs when the input signal goes below zero volts, causing the active device to cut off. The fundamental component of the output current, i_o, is found by apply Fourier series analysis. To illustrate the effect of signal clipping, the sinusoidal amplitude, A, is swept from 0 to 5. The large signal transconductance of this amplifier is plotted in Figure 5.2.



$$\text{Large signal transconductance } G_m = \frac{\text{Fundamental component of } i_o}{A}$$

Figure 5.1: Large signal transconductance of a class AB amplifier

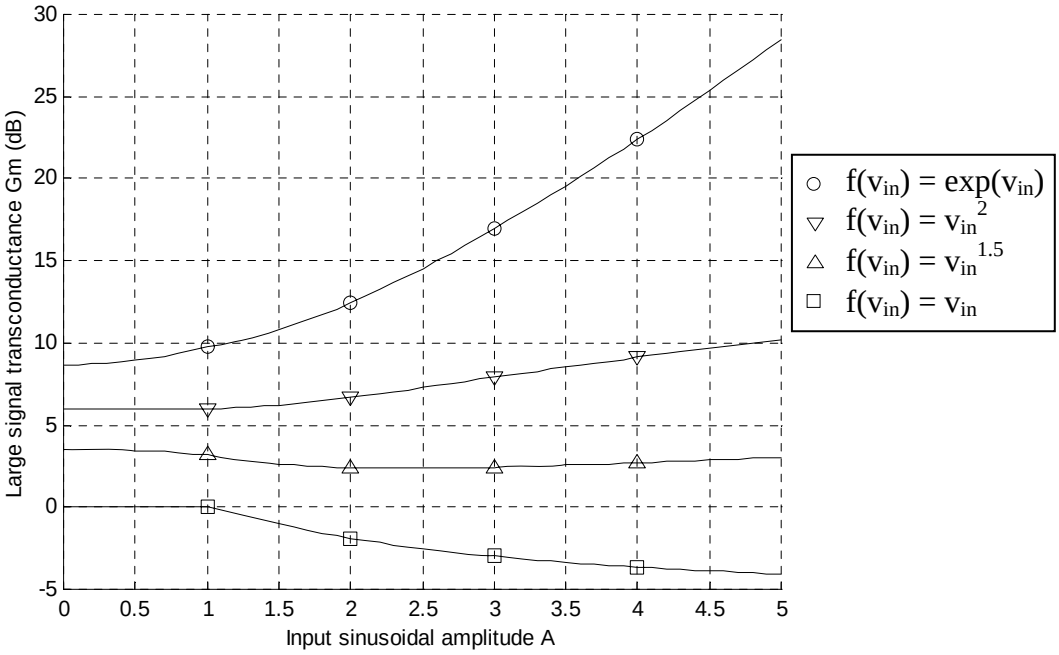


Figure 5.2: Large signal transconductance of different device characteristics

In Figure 5.2, different active device I-V characteristics, $f(v_{in})$, are investigated. When A is less than 1, the amplifier is in class A mode. With linear and squared $f(v_{in})$, G_m's during class A operation are linear as expected. But for $f(v_{in})$ between linear and squared ($f(v_{in})=v_{in}^{1.5}$), G_m compresses. As the amplitude A increases further, the amplifier enters class AB mode. With linear $f(v_{in})$, G_m compresses as expected. However, as for squared $f(v_{in})$, the transconductance expands. G_m for the case when $f(v_{in})=v_{in}^{1.5}$ also shows expansion at large A. One conclusion to be drawn here is that if the active device has a strong enough transfer characteristic, it can potentially have transconductance expansion in class AB mode, even though it has a compression characteristic in class A mode. Note that the exponential I-V characteristic is so strong that it gives transconductance expansion in both class A and class AB modes.

For transistors in the deep submicron CMOS process, the device characteristic barely follows the square law. However, simulations show that when a transistor is properly biased, it can give transconductance expansion behavior. Figure 5.3 shows the transconductance plot obtained by using an NMOS transistor with W/L of 1000μm/0.13μm. The input bias voltage is swept from V_t+10mV to V_t+70mV and its drain is biased at 0.8V by an ideal voltage source.

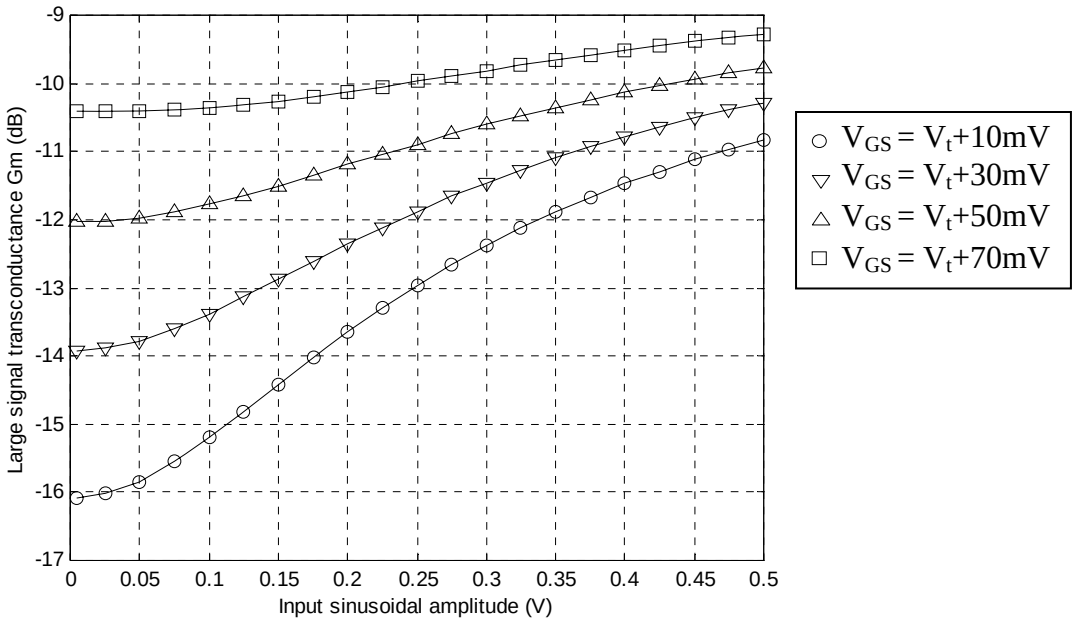
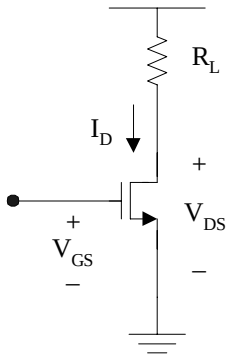


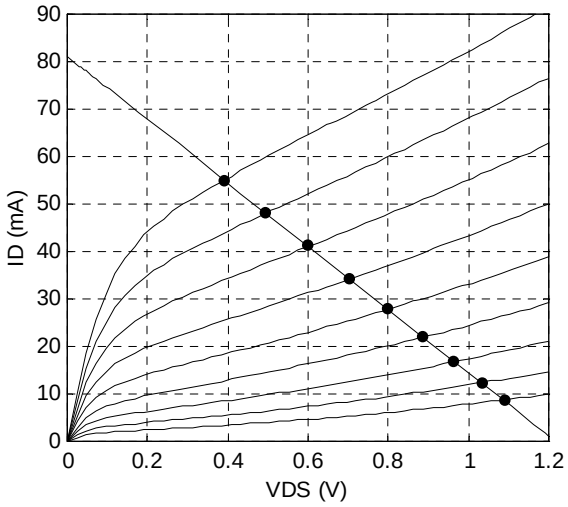
Figure 5.3: Large signal transconductance with NMOS as the active device

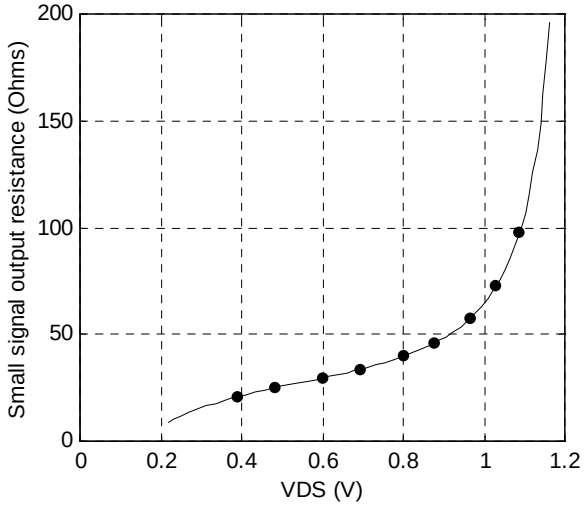
In order to get a linear overall transfer characteristic, a mechanism that causes gain compression must be used to cancel the effect of transconductance expansion. It appears that the nonlinearity of the output resistance is a good candidate for this task. There are three mechanisms that contribute to the nonlinearities of the output resistance: channel length modulation (CLM), drain-induced barrier lowering (DIBL), and substrate current body effect (SCBE). CLM is the result of the effective channel length's being shortened as the device goes into saturation. The output resistance due to CLM is approximately proportional to V_{DS} and it is the dominant mechanism at low V_{DS}.



W/L = 1000μm/0.13μm
R_L = 15Ω

a)





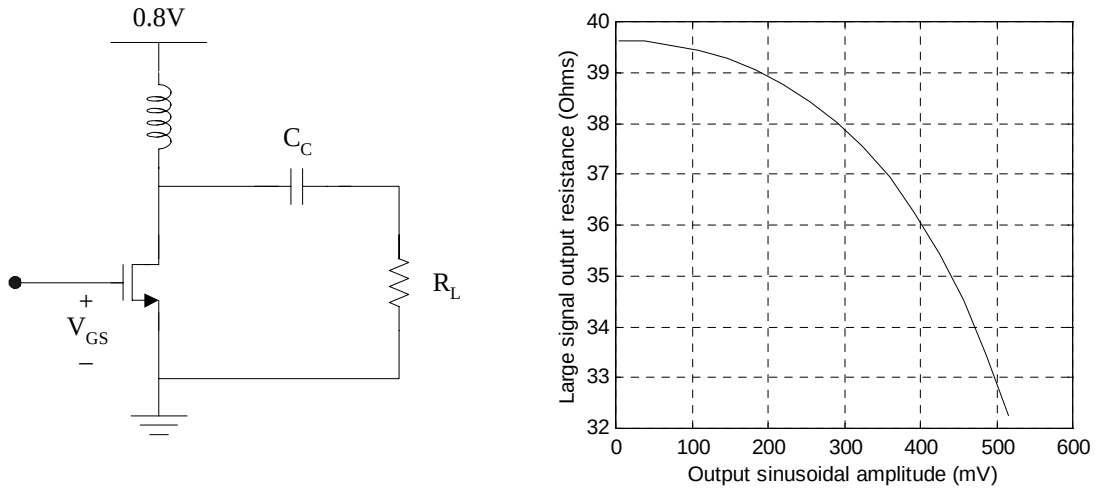
b) c)

Figure 5.4: a) A common source amplifier, b) I_D-V_{DS} and load line plot,
c) Small signal output resistance at each point on the load line

DIBL causes the threshold voltage, V_t, to decrease as drain voltage V_{DS} increases. As V_{DS} increases further, the electric field in the channel can be so high that electrons are greatly accelerated and lead to impact ionization, creating electron-hole pairs (EHP) around the drain region. The electron of an EHP is then swept by the electric field to the drain, causing the drain current to increase,

while the hole is picked up by the substrate, giving rise to additional substrate current. This is known as the hot electron effect. Since the substrate has finite resistance, the substrate current flow causes the body voltage to increase. SCBE is caused by an increase in threshold voltage as a result of this extra substrate current. The output resistance due to SCBE is found to be somewhat inversely proportional to V_D. SCBE is the dominant mechanism at high V_{DS}.

A common source amplifier is considered in Figure 5.4a. The transistor is biased in class A mode by setting V_{GS} at V_t+100mV. The load resistance is chosen to be 15Ω. The load line is superimposed on the transistor I_D-V_{DS} curves as shown in Figure 5.4b. The small signal output resistance of each point on this load line is plotted in Figure 5.4c. At low V_{DS}, the output resistance drops dramatically because of the high current level and the effect of CLM, as discussed earlier. Since the output resistance of a short channel device is poor to begin with, a drop in output resistance can affect the load current considerably, resulting in gain compression. For the circuit used in Figure 5.4a, the small signal output resistance at V_{DS}=0.4V is only 22Ω, which is quite small compared to the 15Ω load resistance. When using this transistor as a class A RF amplifier, the large signal output resistance can be plotted as shown in Figure 5.5b.



a) b)

Figure 5.5: a) An RF common source amplifier, b) Large signal output resistance

Note that this gain compression effect becomes much more significant at high output swing. This effect is unavoidable and quite serious in deep submicron CMOS. It is one of the main reasons why a class A amplifier with a short-channel device has to back off from its peak power considerably in order to maintain good linearity.

With the two mechanisms that create gain compression and expansion discussed above, the task now is to combine both effects to obtain a linear amplifier. A design strategy that can be devised from Figure 5.3 and 5.5b is to allow gain expansion due to device transconductance to start at low output swing and let gain compression due to output resistance take effect at midrange and high output swing. By carefully balancing these two effects, the gain compression

should commence before the transconductance expands excessively, yielding a somewhat linear overall gain characteristic.

As a design starting point, an NMOS with an arbitrary W/L is chosen. The design goal here is to choose an input bias voltage (V_{GS}), input voltage swing ($V_{in,max}$), output voltage swing ($V_{o,max}$) and load resistance (R_L) such that the amplifier is linear while still achieving a reasonably high efficiency. Once the optimal point is found, how much current this transistor can deliver will be determined. From then on, the transistor width and load resistance can be scaled to give the desired output power. Among the four parameters mentioned above, there are three degrees of freedom. One constraint that must be imposed for a linear amplifier design is how much the voltage gain is allowed to vary across the whole linear output range. Even though it is difficult to relate this gain variation (ΔA_v) to linearity specifications commonly listed in wireless standards such as ACPR or EVM, it can at least be an indication of how linear an amplifier is. ΔA_v of 0.5dB or 1dB can be used as an initial design target. Simulations with modulated signal must be done later to verify the actual amplifier linearity. It may take a few iterations before an optimal design can be reached.

To illustrate the design procedure, W/L and ΔA_v of 1000μm/0.13μm and 1dB are used, respectively. Periodic steady-state simulations with different values of V_{GS} and R_L are carried out. For each value of V_{GS} and R_L, the input sinusoidal amplitude, V_{in}, is swept from zero upward until ΔA_v of 1dB is reached. Inductor loss at the output node and power consumption in the driving stage must also be taken into account when computing the overall efficiency. It was

assumed that an inductor with a Q of 20 is used at the output node. As for the power in the driving stage, it was assumed that an inductor with a Q of 5 is used to tune out the input capacitance of the amplifier. RF input power is dissipated mainly in the spiral inductor as the parallel input resistance at the gate is much larger than that of the inductor. Once the RF input power is obtained, it was also assumed that the efficiency of the driving stage is 10%. The maximum overall efficiencies for different values of V_{GS} and R_L are plotted in Figure 5.6.

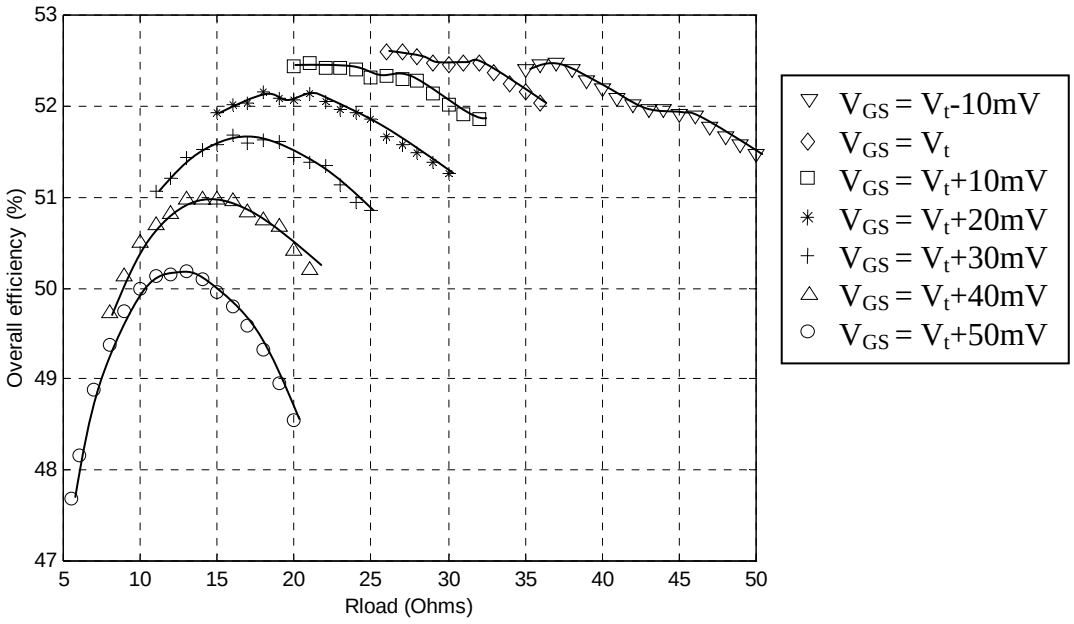


Figure 5.6: Overall efficiency for different values of R_L and V_{GS}

Before a conclusion can be drawn, it is important to understand the simulation results shown in Figure 5.6. From left to right, V_{GS} decreases from V_t+50mV (in class AB operation) to V_t-10mV (in shallow class C operation). We expect to see the overall efficiency trend going up as the bias point decreases

toward class C operation. However, at low V_{GS}, this trend starts to reverse as a result of an increase in RF input drive power, which calls for more power consumption in the driving stage. Furthermore, as V_{GS} decreases, the RL that the amplifier can drive increases and becomes more like the parallel resistance of the inductor, making the loss in the output inductor more prevalent.

For each of these curves, the amount of gain expansion due to device transconductance decreases from left to right. This is because as R_L increases, the gain compression occurs too soon and subdues the effect of transconductance expansion, causing the overall efficiency to be non-optimal. For $V_{GS}>V_t$ cases, the amplifier moves toward class A operation as R_L increases. On some of the curves, especially when V_{GS} is high, the overall efficiency decreases at low R_L. This is because when R_L is low, larger input drive is required in order to get high output swing. Recall that an NMOS transistor enters the triode region when (assuming the transistor follows the square law and does not reach velocity saturation)

$$v_{ds} < V_{DSAT} = v_{gs} - V_T \tag{5.3}$$

With larger input voltage swing, the transistor spends more time in the triode region. Once the transistor enters the triode region, it no longer behaves as a transconductor causing the gain to compress prematurely and hence cause degradation in overall linearity and efficiency. Because of this, for a linear CMOS amplifier design it is generally true that having the active device enter the triode region is undesirable. This limitation on the input swing often makes the transistor in the output stage of a linear PA very large.

Now that the simulation results are understood, a suitable V_{GS} and R_L can be determined. From Figure 5.7, $V_{GS}=V_t+10mV$ and $R_L=25\Omega$ are chosen because with small variations in V_{GS} or R_L, the amplifier is still linear with good overall efficiency. Figure 5.7 shows plots of gain and efficiency versus input amplitude for the amplifier. Note that the gain plot shows that the gain expansion and compression are carefully balanced, giving much better overall efficiency than a class A amplifier.

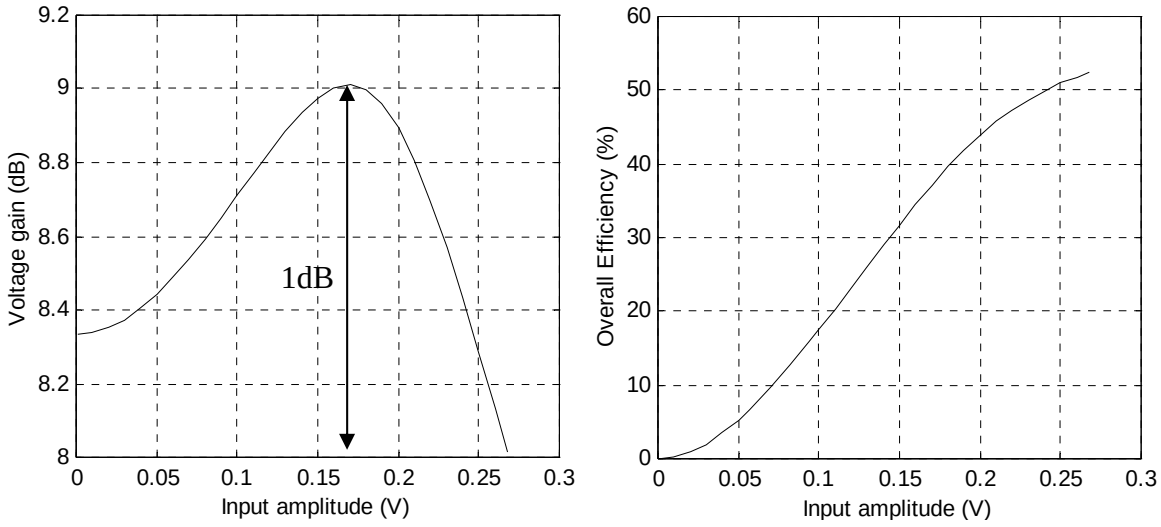


Figure 5.7: Gain and overall efficiency plots with $V_{GS}=0.27V$ and $R_L=20\Omega$

Since the mechanisms that cause gain expansion and compression are not related to each other, it is important to find a biasing scheme that preserves amplifier linearity when subject to process and temperature variations. Since the gain compression due to the active device output resistance is primarily a function of output swing, it is necessary to find a biasing scheme such that the

gain expansion due to the device transconductance is also a function of output swing. For a given load resistance and power supply voltage, the maximum drain current needed to get the maximum output swing ($V_{o,max}$) can be calculated (assuming that $V_{o,max}$ does not change much with process and temperature variations). Therefore, by using a constant current bias scheme, the output swing level at which the transistor enters class AB operation can be controlled. This, in turn, effectively controls the occurrence of gain expansion and compression.

The difference between typical, T=27°C, and fast, T=0°C, corners is small, with the fast corner giving slightly better efficiency. The peak overall efficiency of the slow, T=160°C, corner is a little lower than the other two, mainly due to the temperature increase, but the linearity of the amplifier is still preserved. With the slow corner, the amplifier voltage gain is approximately 1.5dB lower than in the other two cases. This must be compensated for by adjusting the gain of the preceding stage, which should not affect the overall efficiency by much.

One aspect that still requires further research study is the effect of load variation on the linearity. The biasing scheme discussed above is effective when the load resistance is fixed. If the load resistance changes, it may be possible to detect both the output swing and the output power to determine R_L and adjust the bias current accordingly.

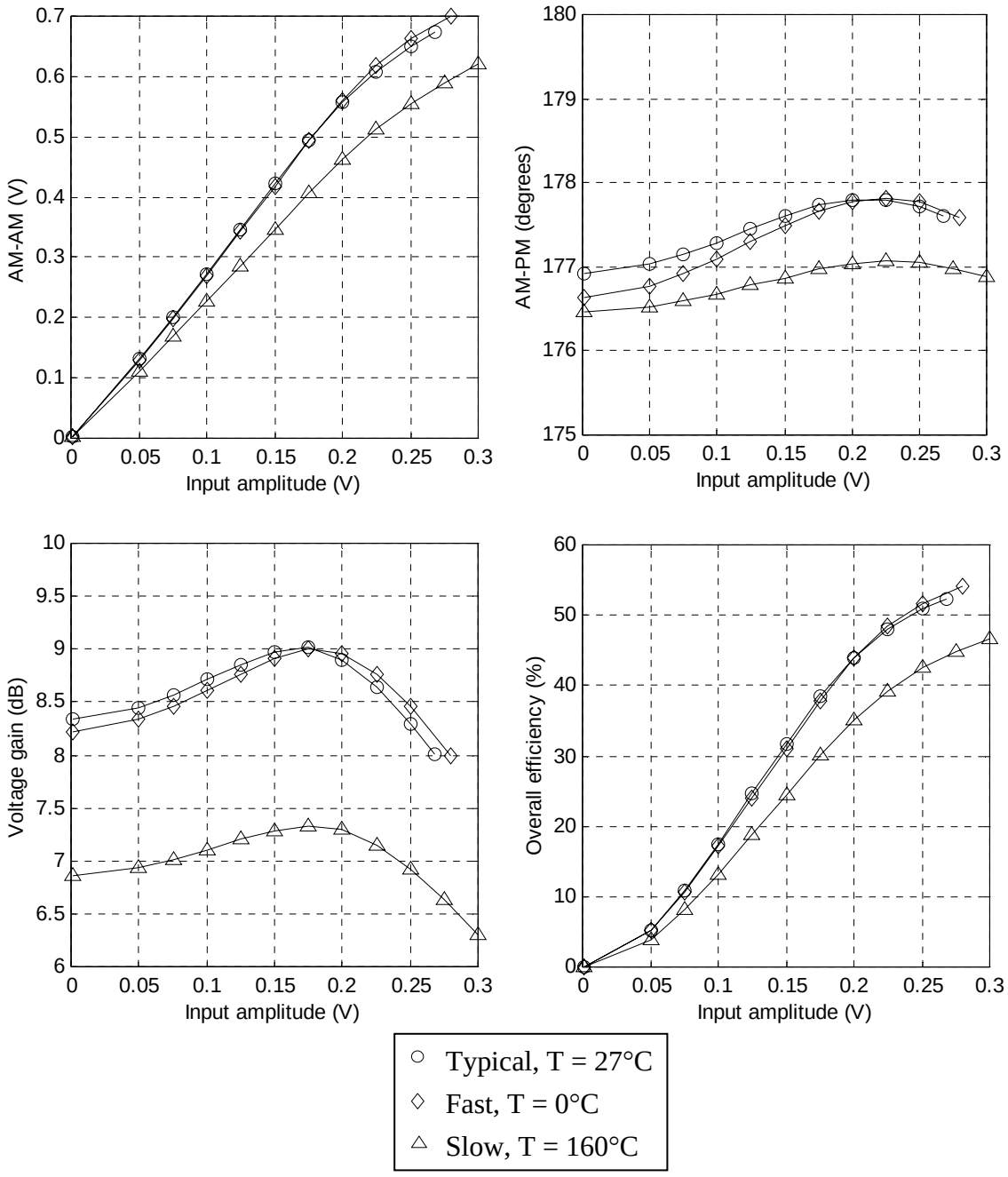


Figure 5.8: AM-AM, AM-PM, voltage gain and overall efficiency at different process and temperature corners

5.2 Linear Output Stage with a Cascode Transistor

For high output power, a cascode transistor is needed to allow the use of higher supply voltage and hence enable higher output swing. The major function of a cascode transistor is to protect the g_m transistor from experiencing destructive oxide breakdown. The transconductance of the amplifier is determined by the g_m transistor, and the impedance looking into the source of the cascode transistor is its load impedance. Therefore, the same design concept can be applied as in a single-transistor common-source amplifier. There are a few tradeoffs in designing a linear amplifier with a cascode transistor that must be taken into consideration. The first tradeoff is in choosing a V_{DD}. Generally, high V_{DD} is desired for better efficiency. For the same output power level, it also reduces the current density, which makes parasitic resistance in interconnects less significant. However, if V_{DD} is too large, the cascode transistor can have a destructive oxide breakdown itself. Larger V_{DD} also means that the voltage swings at cascode transistor terminals vary more, which leads to bigger changes in its parasitic capacitances and higher AM-PM distortion. (C_{db} at the output of a MOS transistor is nonlinear and can create phase distortion.)

The second tradeoff is in choosing the width of the cascode transistor. A large cascode transistor (low V_{DSAT}) allows V_o to go closer to zero, thus yielding higher efficiency. However, this efficiency increase is partially offset by the need for a lower inductance value to tune out its drain-to-bulk capacitance, which in turn leads to more loss in the output inductor. A large cascode transistor also translates to larger parasitic capacitances and therefore higher AM-PM distortion.

Conversely, if the cascode transistor is too small, it can enter the triode region faster and become even more detrimental to the amplifier's overall linearity (both AM-AM and AM-PM). It also causes the load resistance of the g_m transistor to vary rapidly, which degrades the overall AM-AM linearity. Furthermore, once the cascode transistor enters the triode region (assuming the g_m transistor is still in the saturation region), the Miller gain that C_{gd} of the g_m transistor sees expands, causing significant AM-AM and AM-PM distortions in the driving stage. A capacitive neutralization technique can be used to neutralize this effect, as discussed in Section 5.4.

The third tradeoff in designing a cascode transistor is regarding its gate bias voltage, $V_{G,cas}$. $V_{G,cas}$ must be chosen in conjunction with V_{DD} such that the oxide of the cascode transistor does not exceed its breakdown limit. From the linearity perspective, it should be chosen such that both transistors stay in saturation as long as possible. In other words, both transistors should enter the triode region at the same time. However, there is a caveat regarding the order in which these transistors may go into triode. If the cascode transistor enters the triode region before the g_m transistor, it causes the same effects as when the cascode transistor width is too small as discussed in the previous paragraph. However, this is still preferable to the opposite case. If the g_m transistor enters the triode region before the cascode transistor, the g_m transistor stops behaving as a transconductor. Severe AM-AM distortions (in form of gain compression) and AM-PM distortions can occur as a result. In practice, it is nontrivial to design a bias circuit for $V_{G,cas}$ that can achieve the ideal case (where both transistors go

to triode simultaneously) for all operating conditions and process variations. Therefore, the condition where the cascode transistor enters the triode region first is not uncommon.

As a continuance from the amplifier designed in the previous section, a cascode transistor is added to the design and V_{dd} of 3V is used. Cascode transistor width, after a few iterations, is chosen to be the same as that of the g_m transistor (1000μm/0.35μm). The cascode gate is biased with a constant voltage source at 2.2V. A more sophisticated bias circuit can be used for better performance, but is not explored here. The load resistance is also increased to 75Ω (the same proportion as V_{DD} increase). The resulting plots for AM-AM, AM-PM, voltage gain, and overall efficiency are shown in Figure 5.9.

5.3 Driver Stage Design

A typical output stage has 10-20dB of power gain. However, typical analog circuits can deliver less than 1mW, so in order to get to Watt-level output power, one or two driving stages are required. To preserve the overall linearity, the driving stage is usually biased as a class A amplifier. Even though this yields poor power efficiency in the driver itself, it does not substantially affect the overall efficiency since its power consumption is approximately an order of magnitude less than that of the output stage. By using a linear driver, the entire nonlinearity budget can be allocated to the output stage where linearity-efficiency tradeoff is much more severe. On the other hand, a highly inefficient driver can still affect

the overall efficiency. Even though the linearity-efficiency tradeoff in a driver is much less severe, it cannot be overlooked.

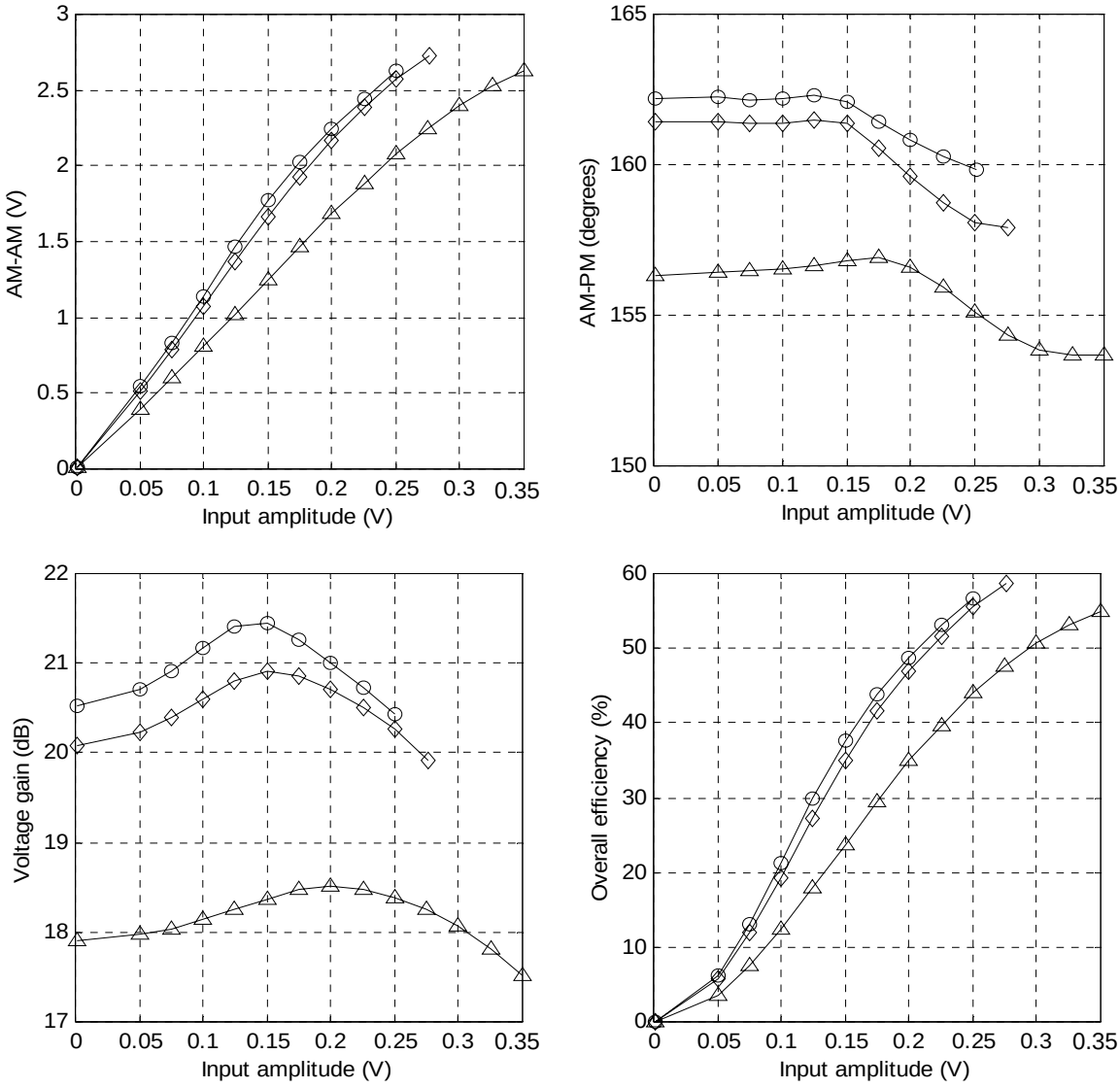


Figure 5.9: AM-AM, AM-PM, voltage gain and overall efficiency at different process and temperature corners

A driver is usually designed as a common source amplifier. A tail current source can be used to improve the common mode rejection ratio if sufficient voltage headroom is available. A cascode transistor may also be used to enable higher supply voltage and higher output resistance, which is very desirable for a class A amplifier. Unlike the output stage, it is not desirable to have the cascode transistor enter the triode region, for linearity reasons. Nonetheless, the capacitive neutralization technique can still be used to improve the linearity of the input impedance as discussed in the previous section. Details of the capacitive neutralization technique will be discussed in Section 5.4.

An inductor is generally used at the output node to tune out the input capacitance of the output stage. This greatly reduces the power consumption but the driver output node now becomes a high-Q node and has limited bandwidth. This is unlike the output node of the output stage, which has a relatively low Q due to small load resistance. When more than one driving stage is used, it is important to carefully align tank frequencies in order to avoid losing voltage gain, which may cause the drivers to be overdriven in order to get the desired output power. Overdriving the drivers is extremely detrimental to the overall linearity. Sizing of a drive stage should be done such that each stage does not give more than 20dB of power gain to avoid instability.

5.4 The Capacitive Neutralization Technique

The load network of a driver stage is usually comprised of the input capacitance (C_{in}) of the output stage, an inductor (usually an on-chip spiral

inductor), and a load resistance that typically is the inductor parasitic resistance. One of the foremost problems in a linear driver design is the nonlinear input capacitance of the output stage, which presents as a part of the driver load. The first source of the problem is when the g_m transistor in the output stage enters the triode region or cutoff region. This causes its C_{gs} and C_{gd} to change significantly. However, this does not pose a serious problem since the g_m transistor does not enter the triode region under normal operating conditions (except perhaps for a very brief instant). Another source of nonlinearity in the input capacitance is when the cascode transistor in the output state enters the triode region while the g_m transistor is still in saturation. This causes the Miller gain across C_{gd} of the g_m transistor to expand, resulting in a higher input capacitance.

For a differential design, cross-coupled capacitors can be used to neutralize this effect, as shown in Figure 5.10. By using this technique, the input capacitance becomes

$$\begin{aligned} C_{in} &= C_{gs} + (1 - A_{vM})C_{gd} + (1 + A_{vM})C_N \\ &= C_{gs} + C_{gd} + C_N + A_{vM}(C_N - C_{gd}) \end{aligned} \tag{5.4}$$

where A_{vM} is the Miller gain, which shows an expansion behavior at high output voltage swing (note that A_{vM} is a negative number). By choosing C_N=C_{gd},

$$C_{in} = C_{gs} + C_{gd} + C_N \tag{5.5}$$

From the equation above, C_{in} is now independent of A_{vM}. Figure 5.11 shows plots of the Miller gain A_{vM} and the large signal C_{in} with and without capacitive neutralization, versus normalized V_{in} of the amplifier designed in the previous

section. Note that the gain drop at high input level is due to the g_m transistor's entering the triode region.

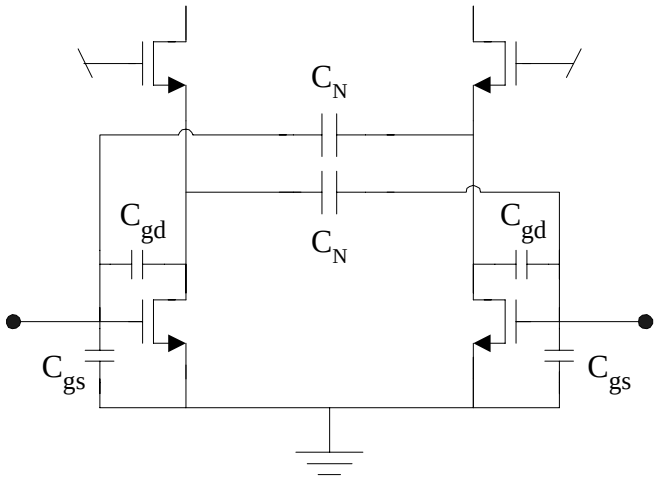


Figure 5.10: Capacitive neutralization

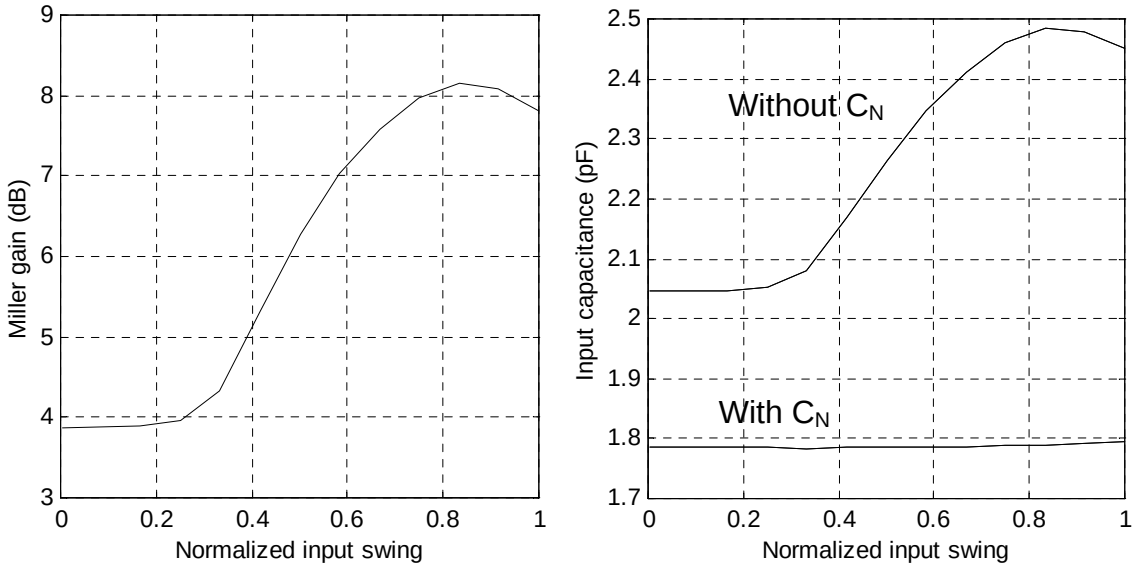


Figure 5.11: Miller gain and large signal C_{in} of a linear output stage

Having cross-coupled capacitors as shown in Figure 5.10 effectively forms a feedback loop which can potentially be unstable. By using a two-port network approach to analyze the loop stability, the circuit from Figure 5.10 can be redrawn as shown in Figure 5.12. Note that L_1 is the effective inductance seen by the input of the output stage.

The loop gain is calculated,

$$Loop\ Gain \qquad LG = \left(\frac{A_{vM}\, s^2\, L_1 C_N}{\frac{s^2}{\omega_0^{\,2}} + \frac{s}{\omega_0 Q_0} + 1} \right)^2 \tag{5.6}$$

where

$$\omega_0 = \frac{1}{\sqrt{L_1 C_T}} \tag{5.7}$$

$$C_T = C_{gs} + (1 - A_{vM})\, C_{gd} + C_N \tag{5.8}$$

and Q_0 is the quality factor of L_1 at ω_0. From Equation 5.6, the term in the parentheses can be re-written as

$$LG = \left(A_{vM}\, s C_N\, Z_{Tank}\right)^2 \tag{5.9}$$

where Z_{Tank} is the impedance of a parallel tank, which consists of $R=\omega_0 L_1 Q_0$, L_1 and C_T. Magnitude and phase of Z_{Tank} are plotted in Figure 5.13. At ω_0, where the magnitude of Z_{Tank} is the highest, the loop gain phase is 180°. For $\omega>\omega_0$, the phase drops and approaches zero at infinite frequency. Therefore, if all assumptions hold, this feedback loop is unconditionally stable since the loop gain phase is never 360°. However in reality, there could be excess phase somewhere in the loop that pushes it into instability. One potential source of the

excess phase is due to A_{vM} not being exactly out of phase from the input. The metric that can be used to determine how tolerant the loop is to excess phase is phase margin (PM).

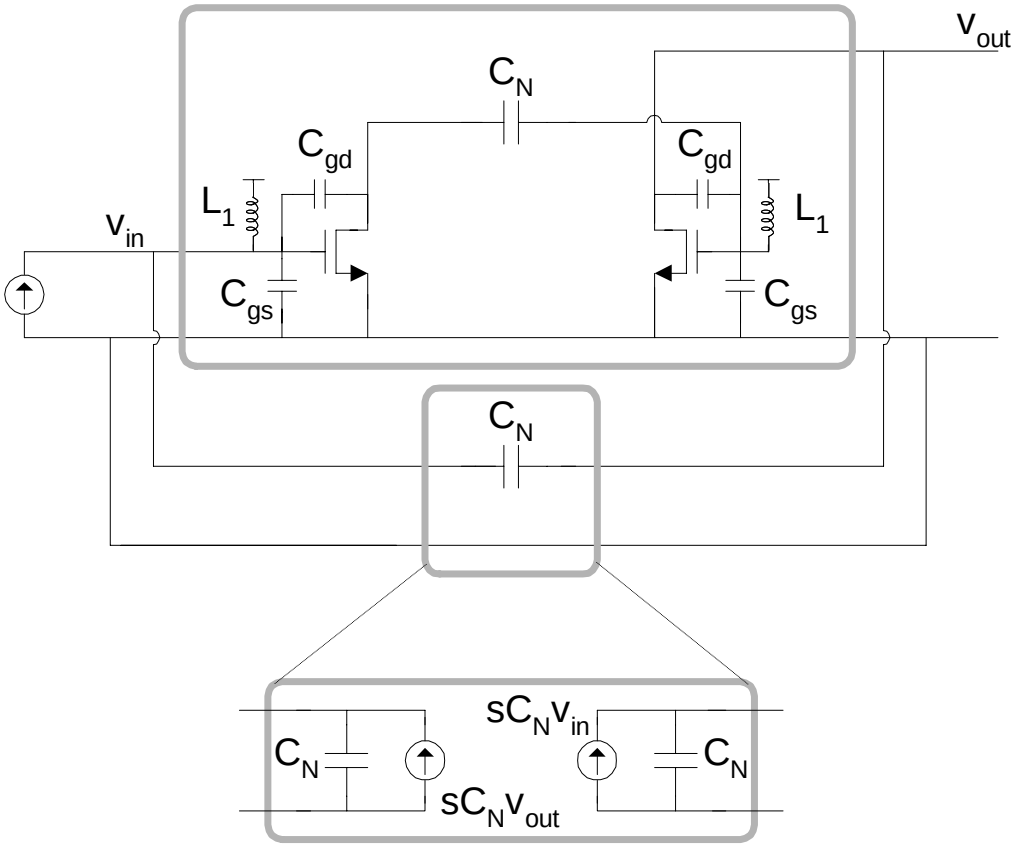


Figure 5.12: Output stage drawn as two port networks for stability analysis

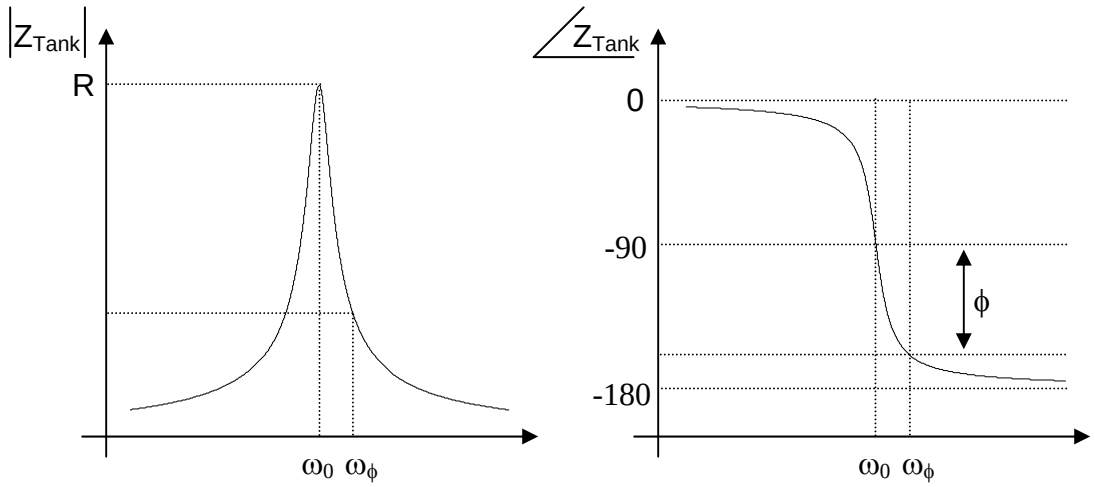


Figure 5.14: Magnitude and phase plot of a parallel tank impedance

According to Figure 5.14, ω_ϕ and the magnitude of Z_{Tank} at ω_ϕ can be found as

$$\omega_\phi = \omega_0 \left(\sqrt{1 + \frac{\tan^2 \phi}{4 Q_0^{\,2}}} \pm \frac{\tan \phi}{2 Q_0} \right) \tag{5.10}$$

$$Z_{Tank}(\omega_\phi) = \frac{jR}{\tan \phi + j} \tag{5.11}$$

$$\Rightarrow \quad \left| Z_{Tank}(\omega_\phi) \right| = \frac{R}{\sqrt{1 + \tan^2 \phi}} = R \cos \phi \tag{5.12}$$

By substituting Equations 5.10 and 5.12 into Equation 5.9 and assuming that the square root term in Equation 5.10 is approximately 1,

$$\left| LG(\omega_\phi) \right| = \left(A_{vM}\, \omega_\phi\, C_N\, R \cos \phi \right)^2 \tag{5.13}$$

$$\left| LG(\omega_\phi) \right| = \left(A_{vM}\, Q_0 \frac{C_N}{C_T} \right)^2 \left(\cos \phi + \frac{1}{2 Q_0} \sin \phi \right)^2 \tag{5.14}$$

To find the phase margin, the magnitude of LG has to be set to 1. Then ϕ and hence phase margin (PM) are found to be

$$\phi = \cos^{-1}\left(\frac{1}{\sqrt{M^2 + N^2}}\right) + \cos^{-1}\left(\frac{M}{\sqrt{M^2 + N^2}}\right) \tag{5.15}$$

$$PM = 180° - 2\phi \tag{5.16}$$

where

$$M = A_{vM}\, Q_0 \frac{C_N}{C_T} \quad \text{and} \quad N = \frac{M}{2Q_0} \tag{5.17}$$

For example, if $A_{vM} = 3$, $Q_0 = 5$ and $C_N/C_T = 0.1$, the phase margin is 71.7°.

One interesting observation is that without the presence of L_1, the loop gain can then be written as

$$LG = \left(A_{vM} \frac{C_N}{C_T}\right)^2 \tag{5.18}$$

which is independent of frequency. Even though the phase of loop gain in this case is always zero, indicating positive feedback, the magnitude of loop gain is usually less than one. It can be proven that the loop gain is equal to one when C_N is so large that C_{in} (Equation 5.4) becomes zero.

5.5 Interstage Matching

As discussed earlier, in designing a linear amplifier in CMOS technology, it is undesirable to have transistors enter the triode region (except for the cascode transistor in the output stage which may enter the triode region slightly before the g_m transistor). Consequently, transistor sizes are made relatively large, to lower V_{DSAT}, but they require small input swing. For a very-high-power amplifier, the input capacitance of the output stage can be very large and must be resonated

by a very small inductor, perhaps too small to be practically realized. A capacitive divider matching network can be used as shown in Figure 5.15.

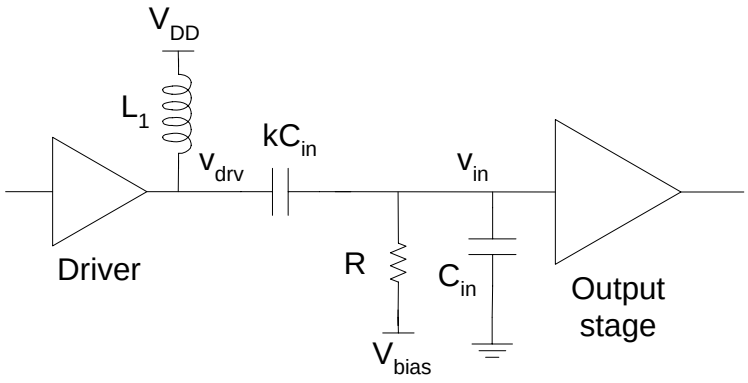


Figure 5.15: Capacitive divider matching network

By placing a capacitor with the value kC_{in} in series with the input, the effective input capacitance is reduced by a factor of k/(1+k). Reduction in the input capacitance can make realization of the on-chip resonating inductor more feasible. The disadvantage of this technique is that the voltage swing (at v_{drv} in Figure 5.15) is increased by the same factor. But since the voltage swing at v_{in} is small to begin with, an increase in voltage swing at v_{drv} can still be delivered by the driving stage. In fact, the maximum output swing of the driver can be used to determine k.

The advantage of this matching network is that it is lossless. Besides, it also eliminates an additional inductor, L_2, at the input of the output stage as shown in Figure 5.16. In Figure 5.16, L_2 is needed to tune out C_{in}. This has to be done right at the input of the output stage in order to avoid large ac coupling capacitor C_c.

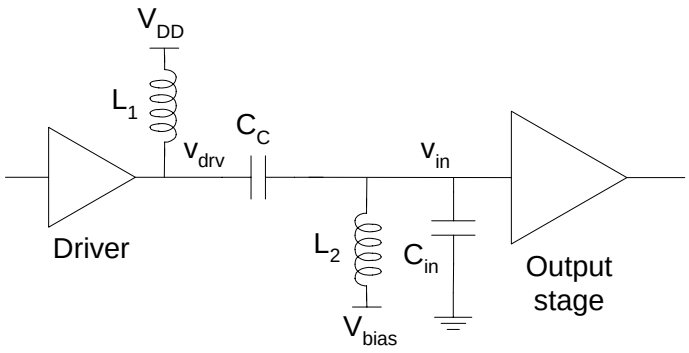


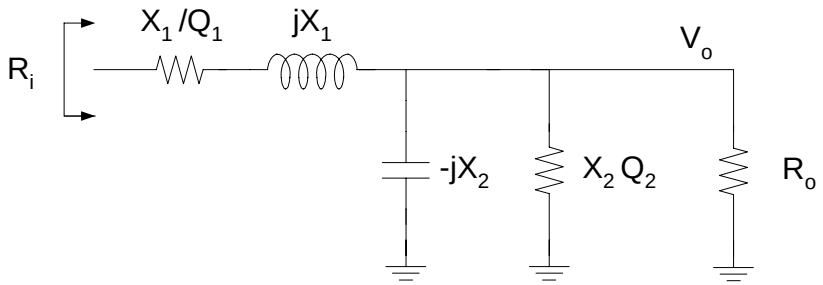
Figure 5.16: A typical interstage matching network

Another advantage of the capacitive divider matching network is that it also improves the stability of the output stage when the capacitive neutralization technique is used. Equations 5.15 to 5.17 show that the stability degrades as the Q of the inductor increases. According to Figure 5.15, the series combination of L_1 and kC_{in}, even though it yields a small inductance, lessens the effective Q of the equivalent inductor by a factor of k/(k+1), bringing better stability of the neutralization loop in the output stage.

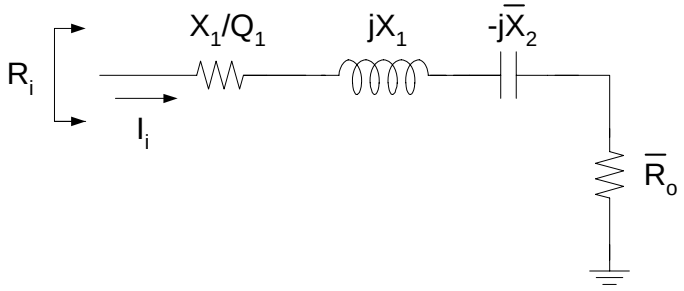
5.7 The Output Matching Network

The main function of an output matching network is to transform the antenna impedance to an appropriate level at the output of the amplifier. All the discussions that follow pertain not only to a linear amplifier design but also to nonlinear amplifier design. This section focuses only on the lumped-element matching network, specifically the L-match topology shown in Figure 5.19a. Readers can refer to [21] regarding the distributed matching network. Instead of

making a high-Q assumption and using approximations when analyzing this network, exact derivations and analysis will be attempted here.



a)



b)

Figure 5.19: An L-match matching network

In Figure 5.19a, it is assumed that both reactive components have finite Q, namely Q_1 and Q_2. For convenience, in the analysis that follows they are modeled by series and parallel combination for inductor and capacitor, respectively. By transforming the parallel branch into a series branch, Figure 5.19b is obtained.

$$\bar{R}_o = \frac{1}{1+Q_{ld}^{\,2}}\left(\frac{X_2 Q_2 R_o}{X_2 Q_2 + R_o}\right) = \frac{X_2 Q_{ld}}{1+Q_{ld}^{\,2}} \qquad (5.19)$$

$$\overline{X}_2 = \frac{X_2 Q_{ld}^{\ 2}}{1 + Q_{ld}^{\ 2}} \tag{5.20}$$

where Q_{ld}, the loaded Q, is given by

$$Q_{ld} = \frac{R_o // X_2 Q_2}{X_2} = \frac{X_2 Q_2 R_o}{X_2 Q_2 + R_o} \tag{5.21}$$

From the above equations, it can be seen that if Q_1 and Q_2 are infinite, the impedance transformation ratio is $1+Q_{ld}^2$ as expected. Now let us define the ideal transformation ratio, C.

$$C = \frac{Desired\ P_{Ro}}{Max.\ P_{Ro}\ without\ impedance\ transformation} \tag{5.22}$$

where P_{Ro} is the power delivered to the load. If the maximum output swing of the amplifier is V_{max}, then

$$C = \frac{Desired\ P_{Ro}}{V_{max}^{\ 2}/(2R_o)} \tag{5.23}$$

For the lossless network case, C is just the transformation ratio $1+Q_{ld}^2$. However, if Q_1 and/or Q_2 are not infinite, C becomes

$$C = \frac{V_{max}^{\ 2}/(2R_i) \times IL}{V_{max}^{\ 2}/(2R_o)} = \frac{R_o}{R_i} \times IL \tag{5.24}$$

where R_i is the input resistance of the network and IL is the insertion loss, which is defined as

$$IL = \frac{Power\ delivered\ to\ the\ load}{Power\ delivered\ to\ the\ load + Power\ loss} \tag{5.25}$$

Both L_1 and C_2 contribute to the network insertion loss. From Figure 5.19a, the insertion loss due to C_2 (IL_2) can be written as

$$IL_2 = \frac{V_o^2 / R_o}{V_o^2 / (R_o // X_2 Q_2)}$$

$$= \frac{X_2 Q_2}{X_2 Q_2 + R_o}$$

$$= 1 - \frac{Q_{ld}}{Q_2} \qquad (5.26)$$

And from Figure 5.19b, the insertion loss due to L_1 (IL_1) can be written as

$$IL_1 = \frac{I_i^2 \overline{R}_o}{I_i^2 (\overline{R}_o + \frac{X_1}{Q_1})} \qquad (5.27)$$

By substituting Equation 5.19 into Equation 5.27 and by letting X_1 in Equation 5.27 be the same as $\overline{X}_2$ in Equation 5.20 (the reactive components must cancel each other),

$$IL_1 = \frac{\frac{X_2 Q_{ld}}{1 + Q_{ld}^2}}{\frac{X_2 Q_{ld}}{1 + Q_{ld}^2} + \frac{X_2 Q_{ld}^2}{Q_1 (1 + Q_{ld}^2)}} = \frac{1}{1 + \frac{Q_{ld}}{Q_1}} \qquad (5.28)$$

The total insertion loss can then be written as

$$IL = IL_1 \times IL_2 = \frac{1 - \frac{Q_{ld}}{Q_2}}{1 + \frac{Q_{ld}}{Q_1}} \qquad (5.29)$$

From Figure 5.19b and Equations 5.19 and 5.20, R_i can be written as

$$R_i = \overline{R}_o + \frac{X_1}{Q_1}$$

$$= \frac{X_2 Q_{ld}}{1 + Q_{ld}^{\ 2}} + \frac{X_2 Q_{ld}^{\ 2}}{Q_1 (1 + Q_{ld}^{\ 2})}$$

$$= \frac{R_o}{1 + Q_{ld}^{\ 2}} \left(1 - \frac{Q_{ld}}{Q_2}\right)\left(1 + \frac{Q_{ld}}{Q_1}\right) \qquad (5.30)$$

Therefore,

$$C = \frac{\left(1 + Q_{ld}^{\ 2}\right)}{\left(1 + \frac{Q_{ld}}{Q_1}\right)^2} \qquad (5.31)$$

With the above results, the design steps can be summarized as follows. Note that the procedure described here can be applied even when L_1 and C_2 are interchanged.

Step 1: Calculate C from Equation 5.23,

$$C = \frac{Desired\ P_{out}}{V_{max}^{\ 2} / (2R_o)}$$

Step 2: Solve for Q_{ld} from Equation 5.31,

$$Q_{ld} = \frac{C\left(1 + \sqrt{1 - \frac{(1 - C)\left(Q_1^{\ 2} - C\right)}{C^2}}\right)}{Q_1\left(1 - \frac{C}{Q_1^{\ 2}}\right)} \qquad (5.32)$$

Step 3: Solve for X_2 from Equation 5.21,

$$X_2 = R_o\left(\frac{1}{Q_{ld}} - \frac{1}{Q_2}\right) \qquad (5.33)$$

Step 4: Solve for X_1,

$$X_1 = X_2 \left(\frac{Q_{ld}^{\ 2}}{1 + Q_{ld}^{\ 2}} \right) \tag{5.34}$$

Step 5: Calculate R_i from Equation 5.30,

$$R_i = \frac{R_o}{1 + Q_{ld}^{\ 2}} \left(1 - \frac{Q_{ld}}{Q_2} \right) \left(1 + \frac{Q_{ld}}{Q_1} \right) \tag{5.35}$$

Step 6: Calculate insertion loss from Equation 5.29,

$$IL = \frac{1 - \frac{Q_{ld}}{Q_2}}{1 + \frac{Q_{ld}}{Q_1}} \tag{5.36}$$

As the impedance transformation ratio increases, a single L-match stage may not be optimal in terms of insertion loss. Multistage L-match may have lower insertion loss if the reactive component or components have low Q, even though the impedance transformation ratio is not high. When designing a multistage L-match, each stage should transform the impedance down by the same factor, i.e., the Q_{ld} should be the same for all stages for optimal bandwidth. With equal Q_{ld} for all stages, a multi-stage L-match network can be designed based on the steps described earlier but replacing C in step 2 with $C^{1/N}$, where N is the number of stages. The component values in each stage can be calculated by repeating steps 3-5 and using R_i in step 5 as R_o for the subsequent stage. The overall insertion loss is then IL^N, where IL is calculated from step 6. To find the optimal number of stages, one can use step 2 (with C replaced by $C^{1/N}$) to calculate Q_{ld} and then use step 6 to calculate IL^N for different values of N and compare. A

multi-stage L-match network also has wider bandwidth compared to a single-stage design, as Q_{ld} of each stage is reduced.

Chapter 6

CMOS Prototype

In order to demonstrate the concept of the Doherty amplifier, a CMOS prototype was designed and fabricated. The goal of this prototype is to achieve over 2W of saturated output power with good efficiency over a wide range of output power. The prototype was designed to be linear enough to meet GSM EDGE linearity requirements in the DCS band, which has the transmit band between 1.710GHz and 1.785GHz. This chapter describes the details of the prototype implementation along with several circuit techniques and practical issues of implementing a highly integrated prototype. The chip was fabricated in a standard digital 0.13μm CMOS process by STMicroelectronics with a MIM capacitor.

6.1 Doherty Amplifier Building Blocks

A block diagram of the implemented Doherty amplifier is shown in Figure 6.1. All the components, including the passive impedance inverter, are integrated on a single CMOS die except for a capacitor used in the output-matching network.

Bondwires are used at the outputs of the output stages and the matching network to achieve high Q inductances.

Each block will now be discussed in detail.

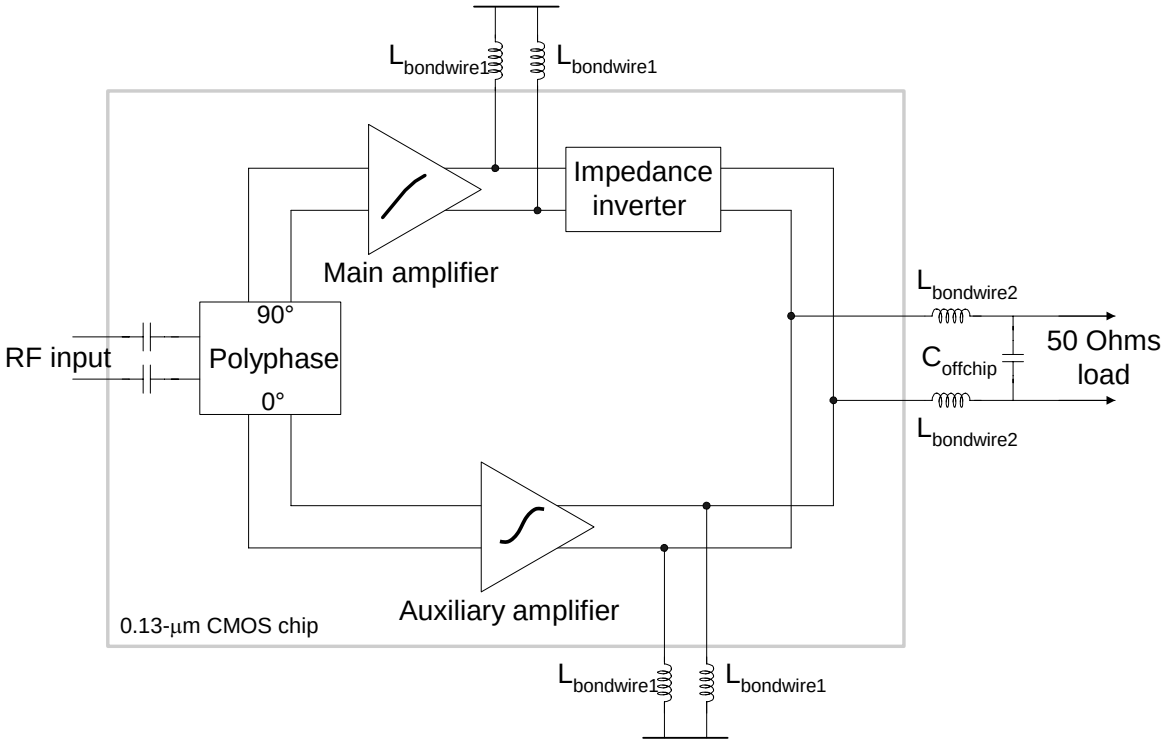
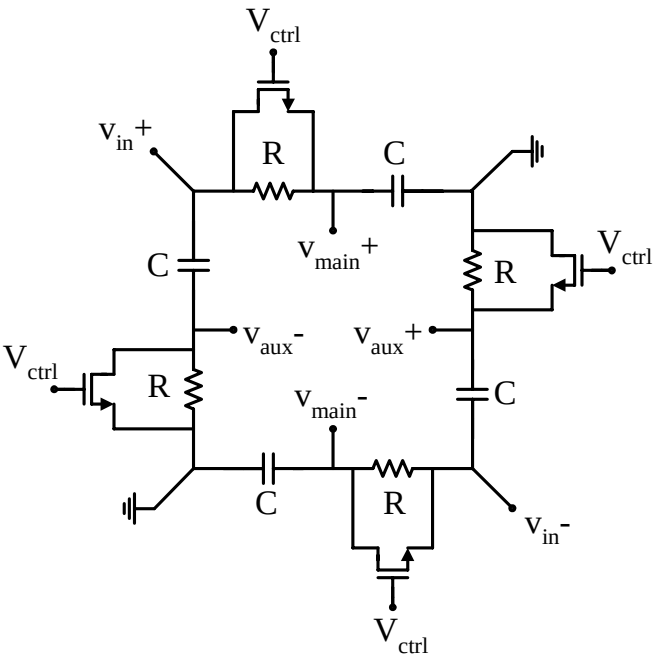


Figure 6.1: Block diagram of the Doherty amplifier prototype

6.1.1 Polyphase Circuit

Doherty amplifier consists of two signal paths, main and auxiliary. The main path has an impedance inverter network, which gives a 90° phase shift, at the output. To equalize the delay of the two signal paths, a polyphase circuit is used as the phase shifter network at the inputs of the two amplifiers. Inserting a phase shift network at the input does not induce extra insertion loss after the amplifier, which would directly lower the overall efficiency. A polyphase circuit is

an RC-CR ladder that gives a 90° phase difference at its two outputs regardless of frequency of operation as long as all R's are matched and all C's are matched. The output phase difference is also insensitive to capacitive load, C_L. A schematic diagram of the polyphase circuit is shown in Figure 6.2.



$$v_{main} = \frac{1}{1 + j\omega R(C + C_L)} v_{in}$$

$$v_{aux} = -\frac{j\omega RC}{1 + j\omega R(C + C_L)} v_{in}$$

$$Relative\ gain = \frac{v_{aux}}{v_{main}} = -j\omega RC$$

Figure 6.2: Polyphase phase splitter schematic diagram

Even though a polyphase circuit can give a robust 90° phase shift, the output amplitudes depend on the magnitude of the impedance of R's and C's, which cannot be matched and therefore is subject to process variations. NMOS resistors are used in parallel with the fixed resistors to compensate for this. Furthermore, they can also be used to adjust the relative amplitude of the two outputs at different output power to get optimal overall efficiency. For this prototype the values of R's and C's are 250Ω and 1pF, respectively. W/L of NMOS transistors is 4μm/0.2μm. Figure 6.3 shows the simulated relative gain of

the two outputs. In order to avoid distortions in NMOS resistors, the input amplitude of the polyphase circuit is kept below 150mV.

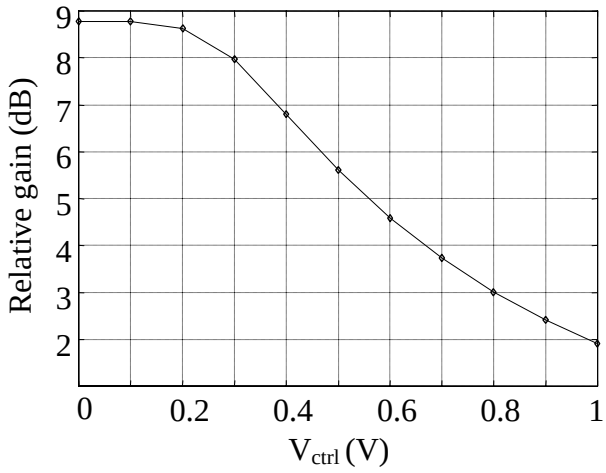


Figure 6.3: Relative gain of polyphase outputs at 1.75GHz

The polarity of the polyphase outputs must be chosen appropriately in order to accommodate the 90° phase lead or lag of the impedance inverter, otherwise the two signal paths will have a 180° phase difference and the two outputs will cancel each other.

6.1.2 Main and Auxiliary Amplifiers

Since the targeted output power is quite high, a driver and a predriver are used in front of the output stage of both amplifiers to ensure that the input power level to the Doherty amplifier chip is small enough that it can be driven by the circuit that precedes it.

6.1.2.1 Output Stage

In an ideal case, the auxiliary amplifier does not affect the overall linearity of the Doherty amplifier. However in reality, due to the finite output impedance of the main amplifier and the finite Q of the Z_{inv} network, nonlinearities in the auxiliary amplifier current can leak to the load, which does affect the overall linearity. As will be discussed later in this chapter, a spiral inductor is used in the Z_{inv} network; it dictates that the auxiliary amplifier still have a certain degree of linearity. In this prototype, the main amplifier is biased as a class AB amplifier for good linearity and the auxiliary amplifier is biased as a shallow class C amplifier for better efficiency without sacrificing too much linearity. In order to simplify the layout of the Doherty amplifier chip, the main and auxiliary are chosen to have exactly the same schematic. This also simplifies the tuning scheme of the Z_{inv} network as discussed in Section 4.6. Even with exactly the same device sizes, the output power of the linear main amplifier is less than that of the auxiliary amplifier. This is because in order for the class AB amplifier to be linear, its input amplitude cannot exceed a few hundred millivolts, whereas the input drive of the class C amplifier can be larger and hence output more power despite a lower bias voltage.

A unit amplifier is first designed in order to determine an optimal biasing point and load impedance. A unit amplifier consists of a g_m transistor with W/L of 1mm/0.13μm and a cascode transistor with W/L of 1mm/0.35μm. The cascode transistor is implemented by a thick oxide transistor in order to allow the use of a 3.3V power supply voltage. After several iterations, it was found that the optimal

point to bias the main amplifier is at V_T+20mV with 180mV input swing and 125Ω as the load impedance for good linearity. As for the auxiliary amplifier, it is optimal to bias at V_T-20mV with 330mV input swing and 70Ω as the load impedance.

The AM-AM and AM-PM characteristics of both amplifiers are shown in Figure 6.4. The peak efficiency is 49% for the main amplifier and 55% for the auxiliary amplifier. Note that the auxiliary amplifier has somewhat higher efficiency while the linearity is not much worse.

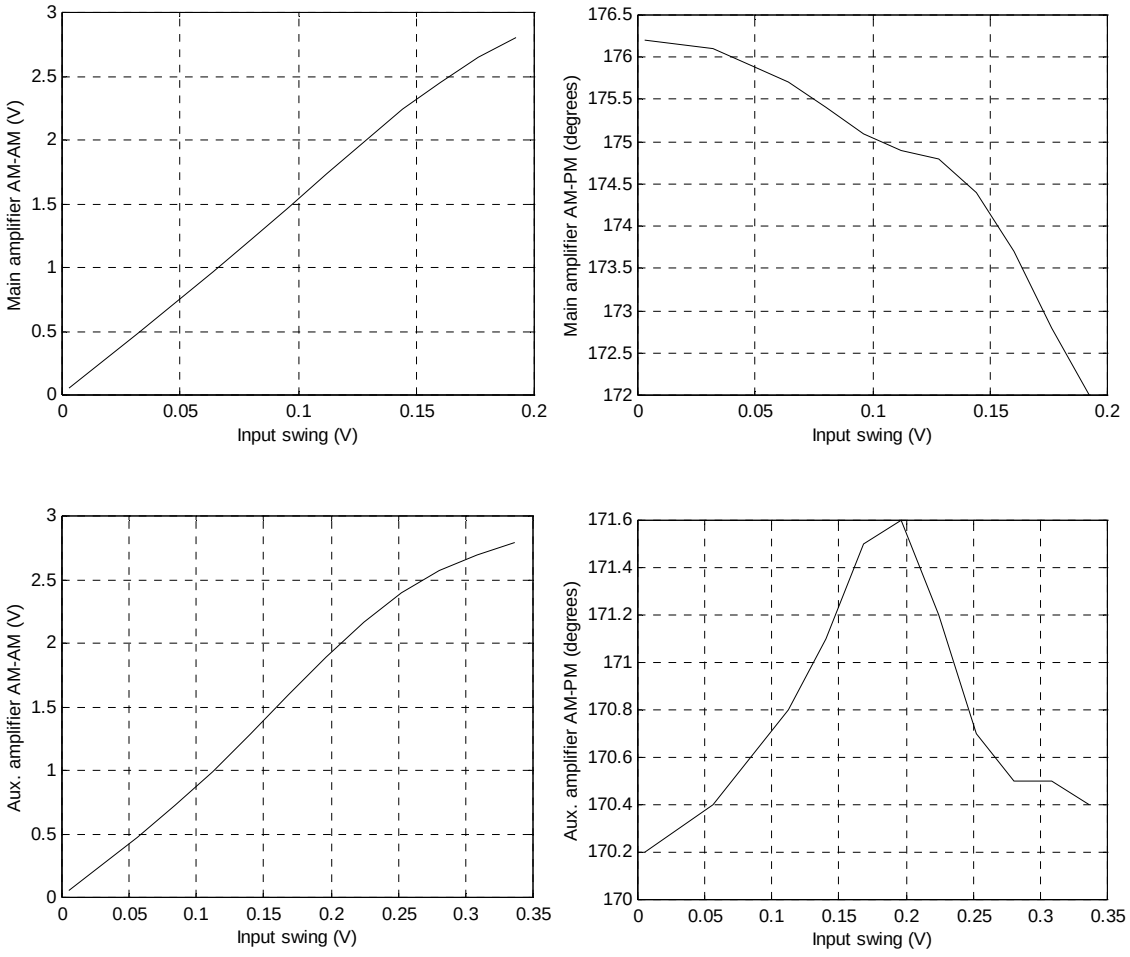


Figure 6.4: AM-AM and AM-PM of the main and auxiliary amplifier

The combined saturated power that these unit amplifiers can deliver is 120mW. Therefore in order to achieve 2W output power, both amplifiers have to be scaled up properly. In the prototype design, a 1.5dB margin in output power is used. As a result, the required scaling factor is 11.75 (since the circuit is differential, the scaling factor is cut by half).

Conventionally, the gate of the cascode transistor is heavily bypassed by a large capacitor to make an AC ground. In this prototype, a different strategy is used and the cascode gate is not completely bypassed, for two reasons. First, due to the large C_{gs} and C_{gd} of the cascode transistor, an excessively large capacitor would be needed in order to completely bypass the AC signal at that node. Besides, by not completely bypassing the cascode gate, C_{gd} of the cascode transistor and the bypass capacitor (C_{bypass}) form a capacitive divider, which allows the voltage swing at cascode gate to be in phase with the output node. This reduces the maximum voltage across the oxide region of the cascode transistor, as shown in Figure 6.5.

From simulations, it was found that C_{gd} of the g_m transistor is approximately 4pF, and the neutralization capacitors, which are implemented by MIM capacitors, are designed accordingly. Even if the neutralization capacitor does not exactly match the value of C_{gd}, a first-order cancellation can still be achieved and improve the linearity significantly.

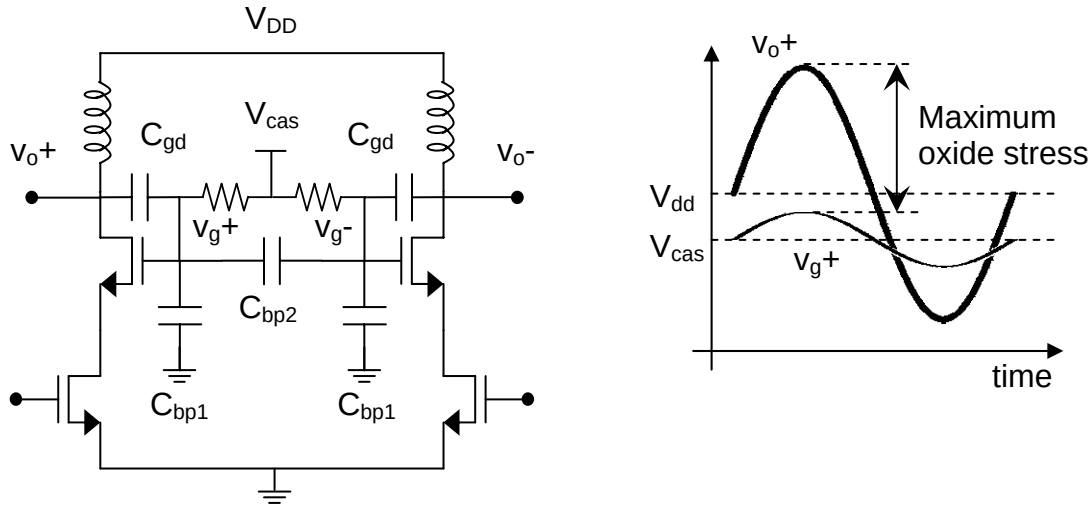


Figure 6.5: Capacitive divider at the gate of the cascode transistor

6.1.2.2 Interstage Matching

The input capacitance of the designed output stage is 21pF. If In order to resonate this input capacitance, a 0.39nH inductor is needed. Capacitive divider matching is used to lower the input capacitance to a more manageable level. How much the input capacitance can be lowered depends on the level of signal swing that the driver stage can provide. Recall that the signal swing is increased by the same factor as the input capacitance reduction (see Section 5.5). Before using this interstage matching technique, it is important to understand its impact on the power consumption of the driving stage. As shown in Figure 6.6, if the effective input capacitance is reduced by a factor of b, the inductance and its parallel parasitic resistance are also increased by a factor of b. Therefore, the power loss in the inductor is increased by the factor b. As a result, the factor b should be made as small as possible in order not to waste too much power in the

spiral inductor, hence lowering the efficiency of the driving stage and the overall efficiency. Another way to look at this is by recognizing that the AC current required from the driver is constant regardless of the value of b. Therefore it is beneficial to keep the power supply voltage of the driver as small as possible. For this reason, a 1.2V power supply is used in this prototype.

With a 1.2V power supply, a common source amplifier stage with a cascode transistor can be carefully designed to have a linear output swing of 400mV. With the 200mV input swing to the main amplifier designed earlier, the series capacitor should then be designed to have approximately the same capacitance value as the input capacitance, which is 21pF, for b=2 (k=1). As for the auxiliary amplifier, using b=2 would require 700mV at the output of the driver. This can cause the driver gain to slightly compress but it does not significantly affect the overall linearity. Using the same value of b for both amplifiers allows them to have exactly the same schematic diagram, which greatly simplifies the layout.

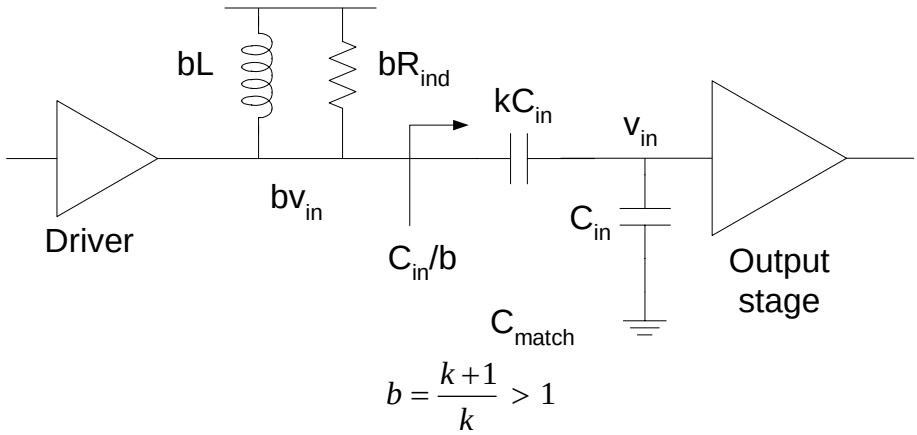


Figure 6.6: Capacitive divider interstage matching

6.1.2.3 Predriver and Driver

The main and auxiliary amplifier output stages designed in the previous section give approximately 23dB and 21dB of power gain, respectively. Two driving stages, predriver and driver, are used in this prototype in order to lower the input drive level and capacitance. Having small input capacitance also prevents excessive loading on the polyphase circuit, which translates to a large amplitude imbalance between the two outputs. Both stages are designed as a common source amplifier with a cascode transistor. A 1.2V power supply is used for both stages.

For the driver, a 0.6nH inductor is used at the output to tune out its parasitic capacitance and the capacitance presented by the output stage and the interstage matching. The designed inductor is estimated to have a Q of 7. (Q was found to be 11 from 3-D electromagnetic simulations in Ansoft HFSS, but a lower number is used in circuit simulations as a safety precaution.) This translates to a 46Ω parallel resistance. For the main amplifier driver, class A operation with 400mV of linear output swing is desired, so the bias current is chosen to be 25mA. As for the auxiliary amplifier driver, since d the desired output swing is a little larger, 35mA bias current is used. In order to get high output swing, the tail current source is omitted in this stage. The g_m and cascode transistors are designed to have W=1500μm and 2800μm, respectively. Note that the channel length used for both transistors is 0.15μm rather than the minimum length, to boost the output resistance for better linearity.

The required input voltage swings for the drivers were found from simulations to be 40mV for the main amplifier and 55mV for the auxiliary amplifier. For simplicity, the same capacitive divider matching strategy with b=2 is used. With b=2, the output swings of the predrivers are merely 80mV and 110mV, which allows the use of a tail current source to improve the common mode rejection of this stage. A 3.1nH inductor is used at the output of the predriver to tune out the parasitic capacitance and the capacitance presented by the driver and the interstage matching. With Q=7, the equivalent parallel resistance is approximately 240Ω. For class A operation, 2mA is used as the bias current in each leg. A 4pF series capacitance is put at the input of the predriver to lower the input capacitance, which is the load of the polyphase, 0.4pF. With the series capacitance, the required input swing is 12mV for the main amplifier and 17mV for the auxiliary amplifier. These voltage swings are still small enough to prevent any distortions from the MOS resistors in the polyphase phase shifter. The g_m and cascode transistors are designed to have W=240μm and 480μm, respectively. The channel length used for both transistors is 0.15μm. As for the tail current source, W/L is 3840μm/1μm.

6.1.3 Output Matching Network

For this prototype, the desired output power is 2W and the power supply in use is 3.3V. Therefore, the required load resistance is

$$P = \frac{1}{2}\frac{V_{\max}{}^2}{R_{Load}} \quad \Rightarrow \quad 2 = \frac{1}{2}\frac{(2x3.3)^2}{R_{Load}} \quad \Rightarrow \quad R_{Load} = 10.9\Omega \qquad (6.1)$$

Note that V_{max} is assumed to be the same value as V_{DD} and the factor of two is due to the amplifier differential output. It is a good practice to overdesign the output power, as there could be a significant amount of power loss from bondwires, external balun, and traces on the printed circuit board. For this prototype, a margin of 1.5dB (a factor of 1.41) is used. That translates to R_{Load} of approximately 8Ω differential or 4Ω single-ended. Note that this margin does not include the loss coming from the output-matching network, as it is already taken into account by the design procedure discussed in the previous chapter. For measurement purposes, an off-chip balun is used to combine the differential outputs. Assuming that the balun has a 50Ω balance port impedance, the ideal transformation ratio required is 50/4 = 12.5.

The output matching network is implemented by using two bondwire inductors, which together with an off-chip capacitor form an L-match section. A spiral inductor is not used here, as it would give a very high insertion loss. To find the optimal number of L-match stages, the procedures discussed in the last chapter are used. The results are summarized in Table 6.1, which shows the insertion loss of the network for different number of stages (n) and several values of bondwire Q (Q_1). The Q of an off-chip capacitor is generally much higher and is conservatively assumed to be 60. Table 6.1 also shows the insertion loss for the case in which the ideal transformation ratio is reduced by half. This can be implemented by using an off-chip balun with 25Ω balance port impedance rather than 50Ω.

Table 6.1: Insertion loss of the matching network with different component Q's

n	c = 50/4			c = 25/4		
	Q1=15 Q2=60	Q1=30 Q2=60	Q1=45 Q2=60	Q1=15 Q2=60	Q1=30 Q2=60	Q1=45 Q2=60
1	1.60	0.88	0.67	1.03	0.58	0.44
2	1.36	0.72	0.52	1.09	0.59	0.43
3	1.52	0.80	0.57	1.31	0.70	0.50
4	1.75	0.91	0.64	1.56	0.82	0.58

From Table 6.1, it can be seen that a two-staged L-match has the lowest insertion loss for most of the cases. However, it has a disadvantage is in that it requires more off-chip components. Since the insertion loss of a single-stage L-match is not much higher, it is chosen over a two-staged design in order to minimize the number of off-chip components. Another reason that a single-stage design is preferable is because it yields more practically realizable component values. Figure 6.7 shows the schematic for both ideal transformation ratios, designed to operate at 1.75GHz. Q_1=15 and Q_2=60 are assumed for both cases.

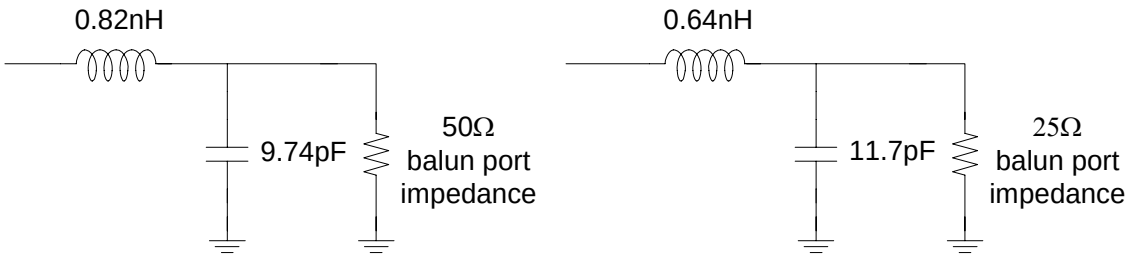


Figure 6.7: L-match networks for c=50/4 and 25/4 at 1.75GHz

The series inductor can be realized by using multiple bond pads in parallel and having multiple bonds per pad. It was found that by having multiple bonds in

parallel, the inductance decreases only marginally due to mutual coupling between bonds. An alternative is to have the die thinned down to minimize the bondwire length. Typically, a die can be thinned down to approximately 200μm. This also improves the PA's thermal characteristic, since the distance between the active circuits and the printed circuit board is shortened. As for the shunt capacitor, the capacitance values can be reduced by half if it is connected differentially. A capacitor with a lower capacitance value generally has higher self-resonant frequency. This may have a significant effect on the component Q since the Q drops significantly as the operating frequency approaches the self-resonant frequency.

6.1.4 The Impedance Inverter Network

The auxiliary amplifier that was designed earlier is capable of outputting 1.85 times more power than the main amplifier. If the impedance inverter network, Z_{inv}, is lossless, its characteristic impedance must be 2.85 times the load resistance. However if the Z_{inv} network is lossy, it affects the output power coming from the main amplifier more than that from the auxiliary amplifier. For proper operation of the Doherty amplifier, this can be compensated for by choosing the characteristic impedance of the Z_{inv} network to be higher—three times the load resistance in this prototype. This factor of three is obtained from iterative simulations.

In Doherty's paper in 1936 [10], a quarter-wavelength transmission was used as the impedance inverter (Z_{inv}). However, a quarter-wavelength

transmission line at 1.75GHz frequency is a few centimeters long and is not suitable for integration. There are many lumped-element networks that can be used in lieu of a bulky transmission line as discussed earlier and as shown in Figure 3.4. In the prototype design, the network in Figure 3.4a is used. The figure is redrawn again here as Figure 6.7 for convenience. The main reason for choosing this topology is its ability to partially absorb the output capacitance (C_{db}) of both amplifiers into the network. Switched capacitor banks are used for a couple of reasons. They are used to match the impedance of all three reactive components in the network. Furthermore, they are used to compensate for variations of bondwire inductors (L_{BW}) at both amplifier outputs, which are used to partially tune out both C_{db}'s. The impedance of these three reactive components is termed network characteristic impedance, as it is equivalent to the characteristic impedance of a quarter-wavelength transmission line.

For the series inductor, a spiral inductor is used instead of a bondwire. This is because both amplifier outputs are on-chip and to use a bondwire here would require a non-standard procedure to bond from an on-chip pad to another and might not be cost-effective in production. One could use two bondwires in series by bonding to off-chip and then come back on-chip, but due to the high targeted output power, the required inductance is too low to be implemented by two bondwires in series.

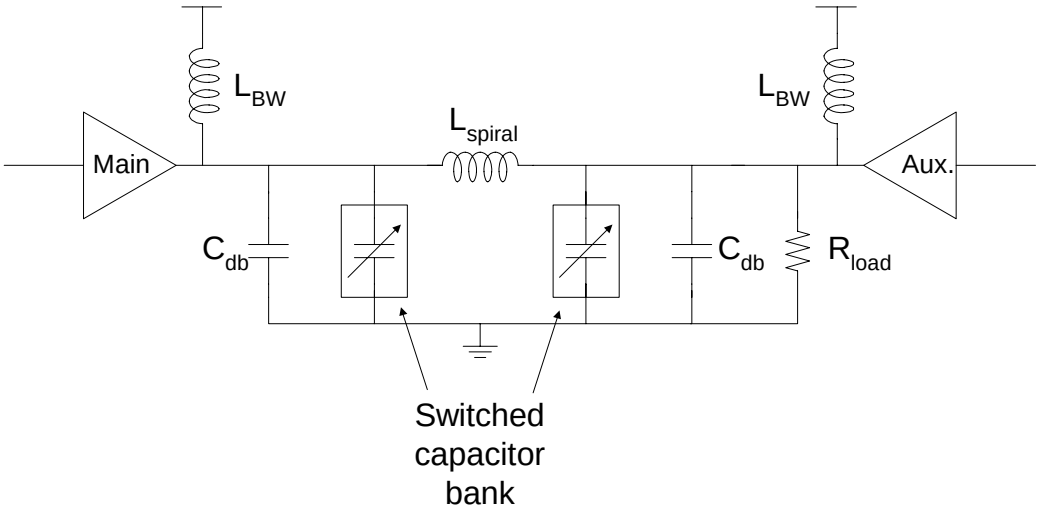


Figure 6.7: Impedance inverter network

Due to the low-Q nature of an on-chip spiral inductor, the power loss in this network causes the overall efficiency to deteriorate. With the main amplifier operating alone, the insertion loss of this network is 1.7dB (Q=7 is assumed). However, once the auxiliary amplifier is turned on, the current from the auxiliary amplifier through the spiral inductor partially cancels that coming from the main amplifier, causing the inductor loss to the output power ratio to decrease. It was found from a simulation that at full output power, the insertion loss of this network is only 0.25dB. Since the loss at the first efficiency peak is significantly higher, it causes the first efficiency peak to be much somewhat lower than the second peak, as will be seen later in the measurement results.

6.1.5 The Switched Capacitor Array

The switched capacitor array used in the Z_{inv} network must be carefully designed to ensure that it has enough range to compensate for process

variations and also for variation of the bondwire inductance. It must also be designed to have low loss. And most importantly, since the array is used at the amplifier output node where signal swing is high, it must be designed such that the voltage across the NMOS switch does not exceed the oxide breakdown voltage.

Figure 6.8 shows a simplified diagram of a switched capacitor array. Note that only one fixed capacitor and one adjustable branch is shown here, as more branches can be replicated and scaled to make a multiple-bit array for coarse and fine tuning. Figure 6.8 also shows a simplified diagram showing states when the switch is open and closed. W in the figure is the width of the NMOS switch in μm. C_{db} and R_{on} are the drain-to-bulk capacitance and the on resistance of a unit transistor, respectively.

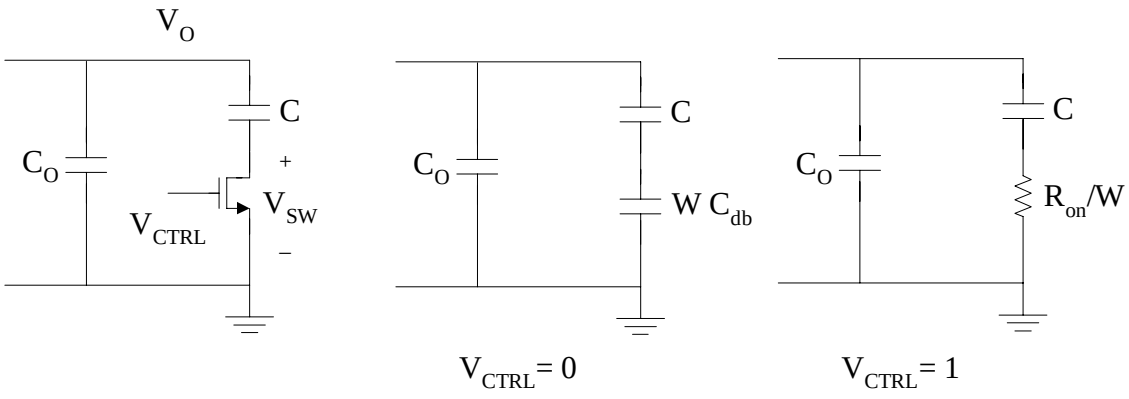


Figure 6.8: A switched capacitor array

In order to design a switched capacitor array, first the minimum and maximum capacitance (C_{min} and C_{max}) of the array must be identified. They can be calculated from the desired frequency range and the process tolerance. The

objective of the design is to achieve the maximum Q without causing any breakdown. Thus, the problem can be formulated as follows. To simplify the calculation, it is assumed that Q>>1.

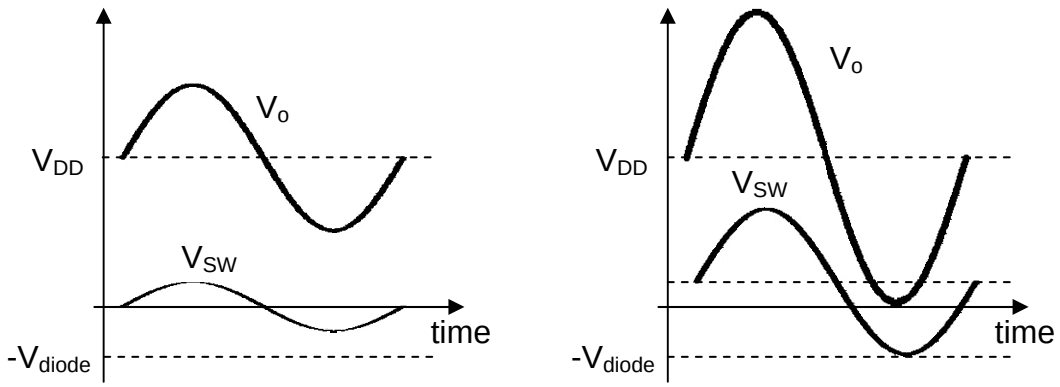
$$\text{Maximize} \qquad Q_{\min} = \omega C_{\max} \frac{R}{W} \frac{1}{1+\left(\omega \frac{R}{W} C\right)^2} \approx \frac{WC_{\max}}{\omega R C^2} \tag{6.2}$$

$$\text{Subject to} \qquad C_{\max} = C_O + C \tag{6.3}$$

$$C_{\min} = C_O + \frac{C C_{db} W}{C + C_{db} W} = C_{\max} - \frac{C^2}{C + C_{db} W} \tag{6.4}$$

$$V_o + 2\frac{C}{C + C_{db} W} - V_{diode} < V_{BD} \tag{6.5}$$

Equation 6.2 is the minimum Q of the array, which is when the switch is on. Equation 6.5 is the constraint on the voltage swing across the NMOS switch, which is problematic when the switch is off (V_{CTRL}=0) with potentially large voltage swing V_{SW}. Figure 6.9 helps explain how Equation 6.5 is obtained. Note the difference in the DC level of V_{SW} with low and high swing at V_o. It is important to recognize that when the switch is off, the DC voltage at the drain of the switch may not be zero if the voltage swing is too large. This is because, with large voltage swing, the drain voltage of the switch can be low enough that the drain to the bulk diode is turned on and supplies extra DC current. In this event, the DC voltage at the drain will increase until the minimum voltage at the drain is higher than the turn-on voltage of the diode, V_{diode}.



Small V_o swing Large V_o swing

Figure 6.9: Voltage swing across the NMOS switch

From the above equations,

$$C = \frac{\Delta + \sqrt{\Delta^2 + 4\Delta C_{db} W}}{2} \tag{6.6}$$

where $\Delta = C_{max} - C_{min}$. Note that the other solution of the quadratic Equation is not valid since it results in negative C. Substituting Equation 6.6 into Equation 6.2 yields

$$Q_{\min} = \frac{4WC_{\max}}{\omega R(\Delta + \sqrt{\Delta^2 + 4\Delta C_{db} W})} \tag{6.7}$$

It can be easily shown that Q_{min} in Equation 6.7 increases monotonically with W. With this result, together with Equation 6.4, it can be concluded that adding a fixed capacitor C_O, despite having very high Q, actually degrades the overall Q of the capacitor array. So the design procedure would be to set $C=C_{max}$, then solve for W from the C_{min} equation (Equation 6.4). If the result does not violate the

inequality in (6.5), the solution is optimal. But if (6.5) is violated, it means that the C_{max}/C_{min} ratio is too large and the array cannot be accommodated without causing oxide breakdown of the switch.

For this prototype, the range of the capacitor array must be able to compensate for 15%, 10%, and 30% variations from the capacitor, spiral inductor, and bondwire, respectively. Assuming that all these factors are independent random variables with Gaussian distribution, the capacitance in the array should thus have roughly a 35% range (the RMS sum of the three numbers above). The desired capacitance at the output node of both amplifiers is governed by the characteristic impedance of the Z_{inv} network. The value of the shunt capacitance can be found to be

$$C = \frac{3R_L}{2\pi f} = 8.66\,pF \tag{6.8}$$

Therefore, C_{min}=5.63pF and C_{max}=11.7pF. However, the problem here is unique in that a fixed capacitance from the amplifier is also presented at the output node. Utilizing the amplifier output capacitance allows the value of the bondwire inductance to be higher, making it more practically realizable and creating lower loss. With that in mind, both C_{min} and C_{max} are reduced, while keeping Δ constant, until the inequality in (6.5) becomes an equality. The array designed using the stated strategy is shown in Figure 6.10. Note that four adjusting branches are used to make a 4-bit array. The achieved adjustment range is $\pm$28%, which is lower than expected due to the presence of the fixed capacitor but is deemed sufficient. The minimum Q of this capacitor structure was found from a simulation to be 55.

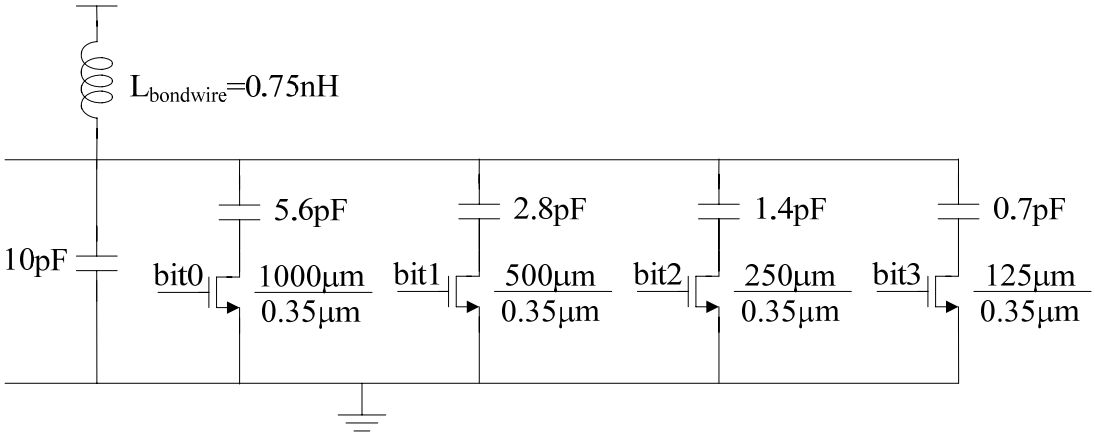


Figure 6.10: The implemented 4-bit switched capacitor array

6.2 Simulation Results

The simulation results for all the circuit blocks designed as described in the previous section are shown in Figure 6.11. These results are obtained by applying a single tone input at 1.75GHz and simulated with periodic steady state (PSS) analysis. Note that the input voltage swing is small enough that it does not create significant distortions in the polyphase circuit. With this input swing and 0.4pF of input capacitance, the prototype easily could be driven by a circuit that precedes it. The overall efficiency plot is obtained by including DC power consumption in all stages. The peak current of the two output stages combined (with a 3.3V power supply) is 1.6A. As for the first two class-A stages, they consume 0.15A from a 1.2V power supply. The input power to the amplifier is not included in the calculation as it is several order of magnitudes below the output power, therefore the overall efficiency can also accurately represent the PAE of the amplifier.

In order to achieve good efficiency at back-off output power, the bias currents of both amplifiers are lowered as the output power decreases. Figure 6.12 shows the overall efficiency obtained at different output power levels. Note that the first efficiency peak is not apparent due to large insertion loss of the impedance inverter network. From this plot, it can be seen that the overall efficiency drops by half at approximately 12dB back-off. This is a major improvement over a conventional stand-alone PA design. For example, a class A amplifier has its drain efficiency reduced by half at only 3dB back-off.

In order to test the linearity of the amplifier, Envelope Following (EF) simulations in Spectre RF were carried out. EF simulation is a time domain simulation which is suitable for RF signals with narrow bandwidth. It takes advantage of the fact that a narrow-banded signal has a slowly varying envelope. Therefore, the circuit transient response of several RF cycles can be skipped since they are similar to the previous cycles. By using GSM/EDGE modulated input signal, the output power of the PA has to be backed-off from the peak power to avoid distortion at high output swing. It was found from simulations that the peak linear output power that the output spectrum still meets the GSM/EDGE mask requirement is +29dBm (5dB back-off). The PAE at this output power level is 31%. With only the main amplifier operating alone, the peak linear output power is +18dBm with 22% PAE.

The EVM of the output signal was simulated in Matlab by using the AM-AM and AM-PM characteristic of the PA. At all the output power levels mentioned

above, the simulated EVM easily meet GSM/EDGE requirements (both RMS and peak EVM).

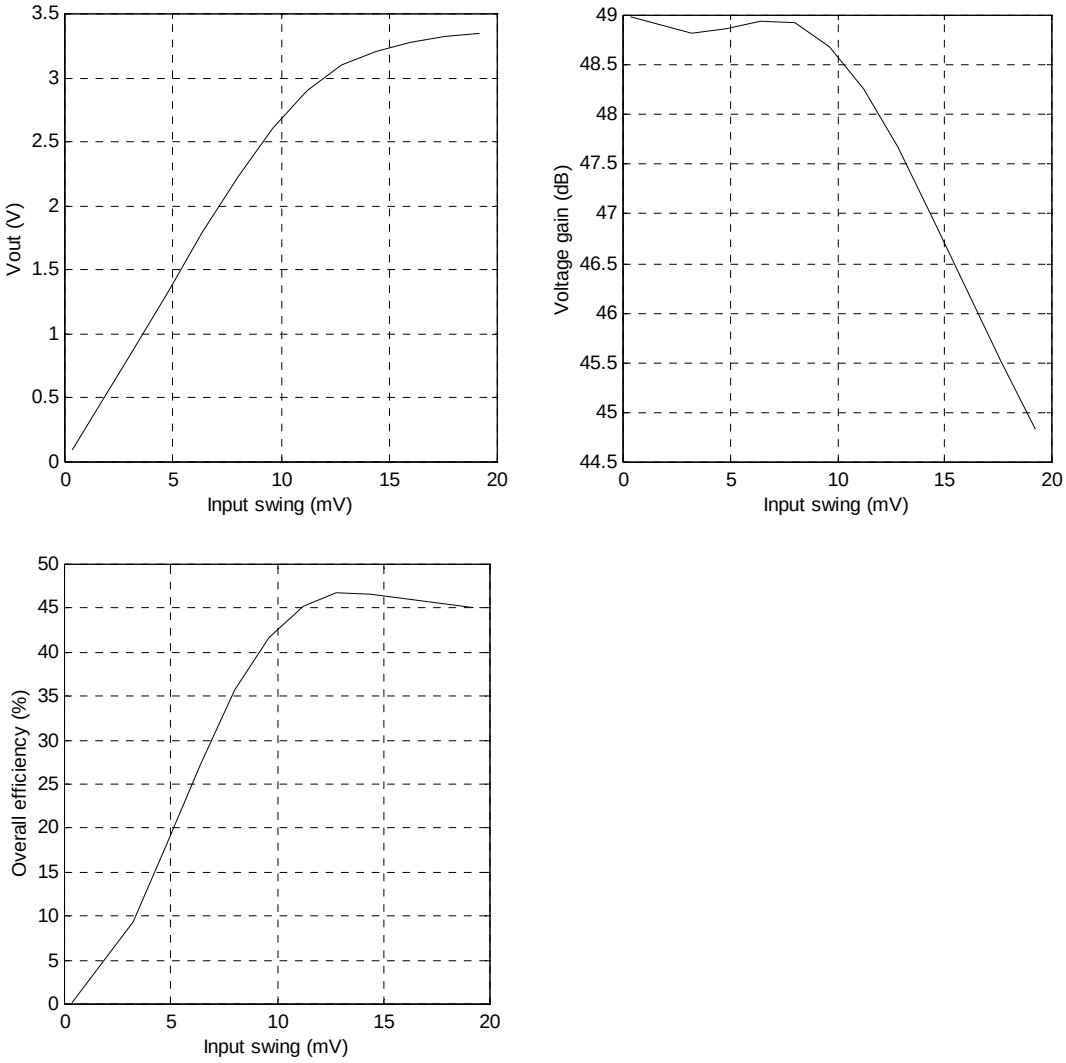


Figure 6.11: PSS simulation results of the prototype

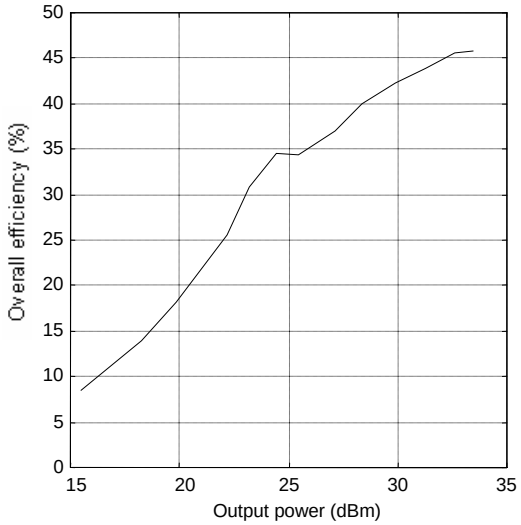


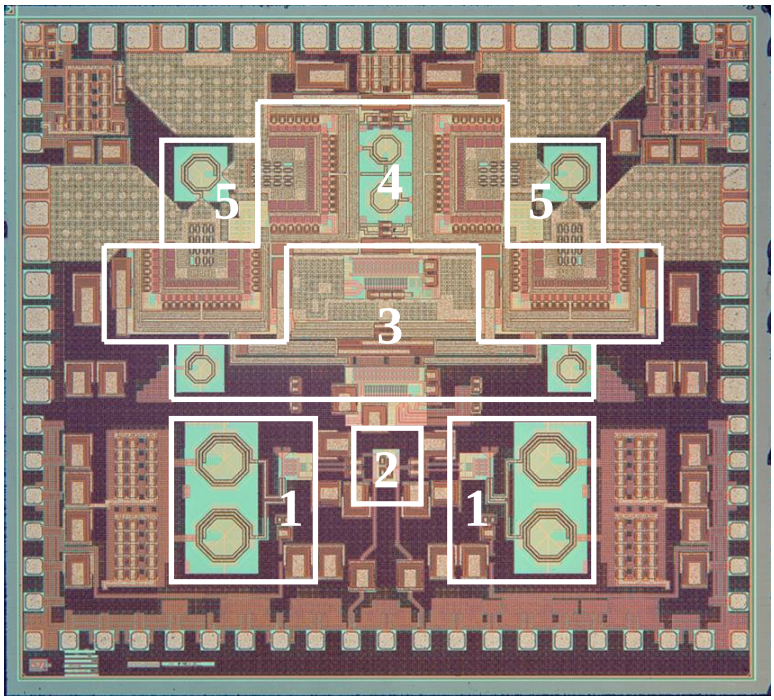
Figure 6.12: Simulated overall efficiency versus output power

6.3 The CMOS Prototype

The prototype was fabricated in a standard digital 0.13μm CMOS process by STMicroelectronics with MIM capacitor. This process is a triple-well process. It is favorable to put the PA in its own well since it has large output swing and may interfere with other circuits. Figure 6.13 shows the chip micrograph.

The die size is 3.2x2.8mm^2 including the pad ring and bypass capacitors. There are 74 pads in total with 2 RF input pads, 4 RF output pads, and 14 DC pads for bias and gain adjustment purposes. The rest are power and ground pads. The RF input signal comes in differentially from the bottom to the polyphase circuit. The RF output signal exits from the top of the die. Note that many pads on the top half of the layout are larger than the others. These pads belong to the output stages where the current level is high. By using larger pads, they can be double-bonded allowing a better current handling capability and

lower inductance value. This is particularly important for the ground inductance as it can generate ground bounce and degrade the performance of the PA.



1. Predrivers
2. Polyphase
3. Drivers
4. Output stages
5. Impedance inverter

Figure 6.13: Chip micrograph

6.4 Measurement Results

The peak output power of 1.5W or approximately 31.8dBm is measured at 1.7GHz. The operating frequency is lower than expected due to excessive bondwire inductance at the outputs of both amplifiers. The peak drain efficiency is measured to be 39%. With the power consumption in the driving stages included, the peak PAE is 36% (33% combined with off-chip balun). The PAE stays above 18% over a 10dB range of output power. Figure 6.14 shows a plot of PAE versus output power. Efficiency of a typical class C amplifier is also plotted for comparison. Assuming the same peak efficiency, the efficiency improvement

of the prototype is significant at large back-off power. For example, at 10dB power back-off, the efficiency of the prototype is almost twice that of the class C amplifier.

When tested with a GMSK modulated signal, the output spectrum fits under the GSM spectral mask at all output power levels. The RMS and peak phase errors are measured to be 0.8° and 1.2°, respectively (the GSM standard required that the RMS and peak phase errors be less than 5° and 20°). Figure 6.15 shows the output spectrum of the prototype. Note that the spectrum clears the mask by over 5dB at 400kHz offset.

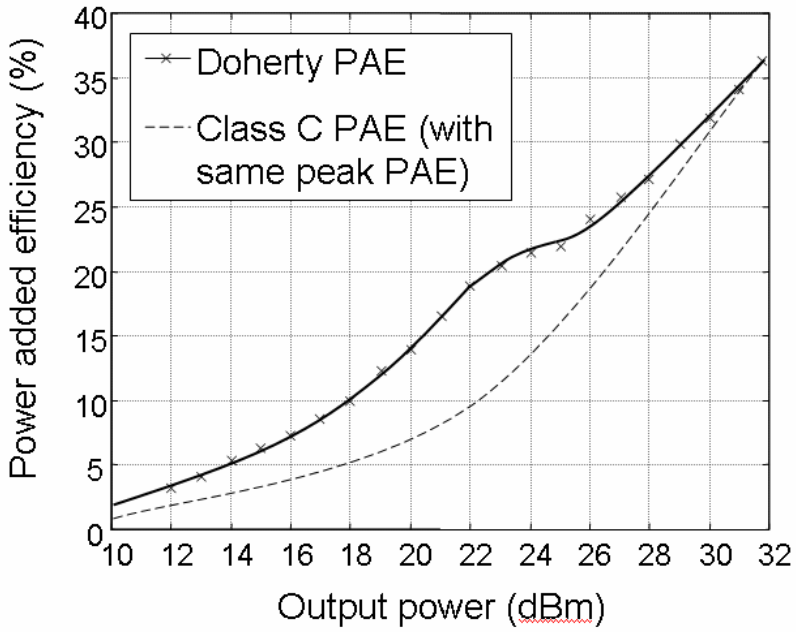


Figure 6.14: PAE versus output power plot

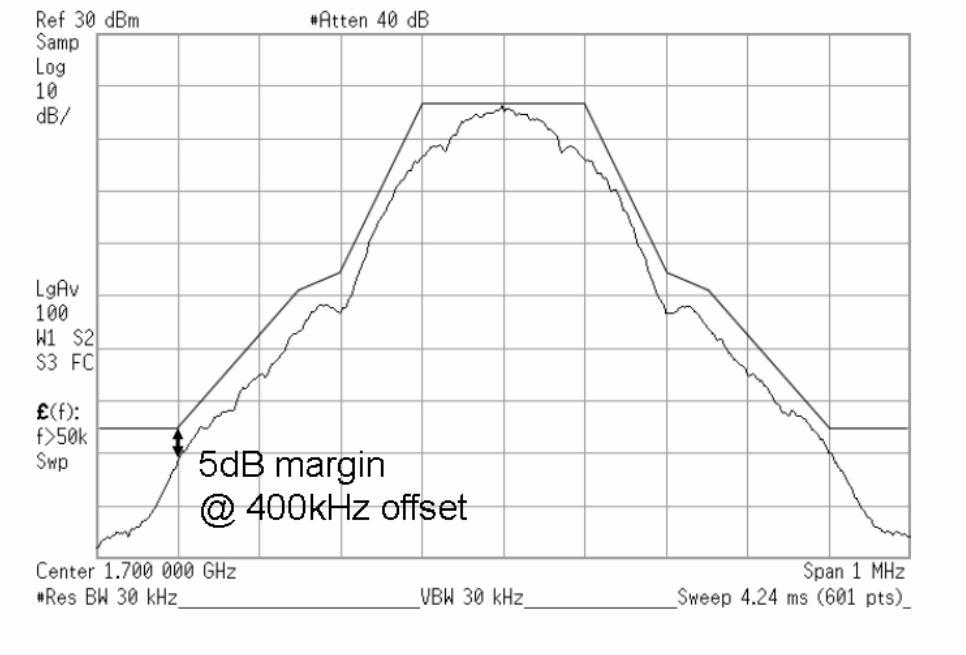


Figure 6.15: Output spectrum with GMSK modulated signal

With GSM/EDGE modulated input signal, the measured peak output power while still meeting the spectral mask requirement is +25dBm with 13% PAE. With the main amplifier operating alone, the linearity requirement is met at +13dBm with 6% PAE. The measured output spectrum at +25dBm is shown in Figure 6.16.

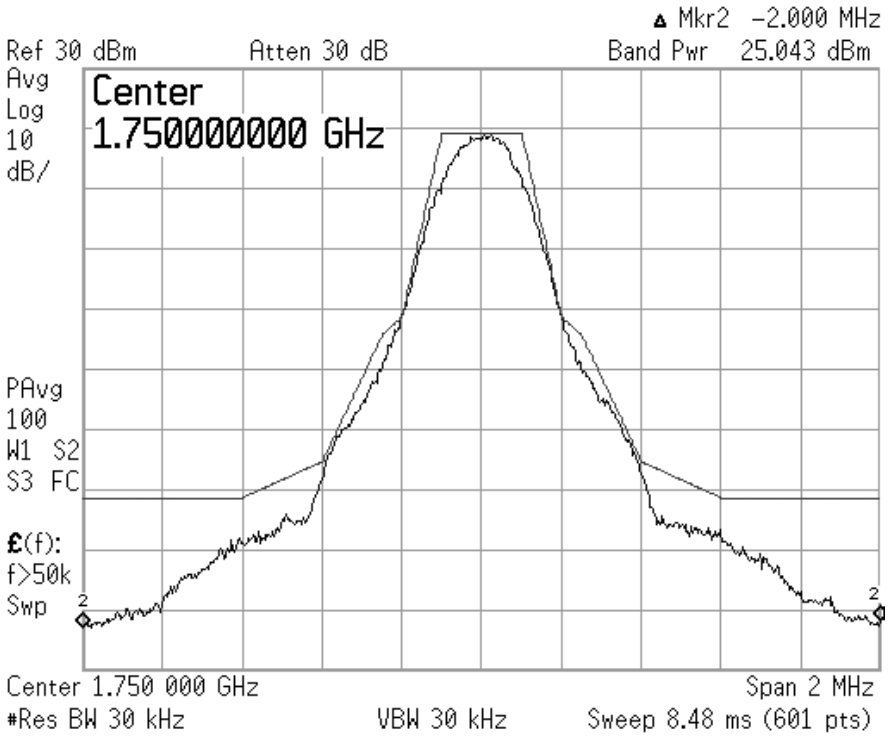


Figure 6.16: Output spectrum with GSM/EDGE modulated signal

Chapter 7

Conclusion

This thesis examines several techniques to enhance power amplifier performance. By applying these techniques to the CMOS PA problem, performance can be improved beyond what can be achieved by a stand-alone CMOS PA. A prototype was built based on the Doherty amplifier technique, which allows an amplifier to achieve high efficiency over a wide range of output power. Various circuit techniques were used to improve the efficiency and linearity and to allow a high level of integration. This work also demonstrates the feasibility of integrating all circuit blocks necessary to implement a Doherty amplifier. Only one off-chip component was required to make the amplifier functional. A lumped-element impedance inverter was implemented by on-chip passive components and is fully integrated. It was shown that when using a low-Q spiral inductor, the power loss in the impedance inverter network does not significantly degrade the peak efficiency. However, with a better inductor, the efficiency at large output power back-off can be greatly improved. A linear class AB amplifier was analyzed, designed, and implemented as the main amplifier.

From simulations, it was concluded that a class AB amplifier can have a linear transfer characteristic with much better efficiency than a class A amplifier. The prototype achieves +31.8dBm peak output power with 36% peak PAE. The efficiency is kept above 18% over 10dB range of output power. The prototype passes the spectral mask and phase error requirements of GSM standard for all output power levels.

Besides the Doherty technique, other enhancement techniques can be used to further improve the PA performance. These techniques include, but are not limited to digital adaptive predistortion, Cartesian feedback, power supply variation, and bias adaptation. With some or all of these techniques working together, a CMOS PA can potentially have performance comparable to discrete PAs. This would enable a SoC integration of high performance wireless transceivers in a low-cost CMOS technology.

References

[1] W.H. Doherty, “A New High Efficiency Power Amplifier for Modulated Waves,” Proc. Institute of Radio Engineers, vol. 24, no. 9, pp. 1163-1182, Sept., 1936.

[2] N. Wongkomet, L. Tee, and P.R. Gray, “A 1.7GHz 1.5W CMOS RF Doherty Power Amplifier for Wireless Communications,” Dig. Tech. Papers, Int. Solid-State Circuit Conf., pp. 486-487, Feb., 2006.

[3] M. Iwamoto, A. Williams, P.F. Chen, A. Metzger, L.E. Larson, and P.M. Asbeck, “An Extended Doherty Amplifier with High Efficiency over a Wide Power Range,” IEEE Trans. on Microwave Theory and Techniques, vol. 49, no. 12, pp. 2472-2479, Dec., 2001.

[4] S.C. Cripps, “RF Power Amplifier for Wireless Communications”, Norwood, MA: Artech House Incorporated, April 1999.

[5] P.R. Gray and R. Meyer, “Future Directions of Silicon ICs for RF Personal Communications,” Proc. IEEE Custom Integrated Circuits Conf., pp. 83-90, May, 1995.

[6] A.A. Abidi, A. Rofougaran, G. Chang, J. Rael, J. Chang, M. Rofougaran, and P. Chang, “The Future of CMOS Wireless Transceivers,” Dig. Tech. Papers, Int. Solid-State Circuit Conf., pp. 118-119, Feb., 1997.

[7] P. Bonnaud, M. Hammes, A. Hanke, J. Kissing, R. Koch, E. Labarre, and C. Schwoerer, “A Fully Integrated SoC for GSM/GPRC in 0.13μm

CMOS," Dig. Tech. Papers, Int. Solid-State Circuit Conf., pp. 482-483, Feb., 2006.

[8] M. Zargari, M. Terrovitis, S.H.M. Jen, B.J. Kaczynski, M. Lee, M.P. Mach, S.S. Mehta, S. Mendis, K. Onodera, H. Samavati, W.W. Si, K. Singh, A. Tabatabaei, D. Weber, D.K. Su, and B.A. Wooley, "A Single-Chip Dual-Band Tri-Mode CMOS Transceiver for 802.11a/b/g Wireless LAN," IEEE J. Solid-State Circuits, vol 39, no. 12, pp. 2239-2249, Dec., 2004.

[9] Lee, H. Xie, L. Wei, L. Luong, J. Pan, S.T. Yang, W.F.A. Lau, W.L. Ngai, "Design of a Low-Cost Integrated 0.25μm CMOS Bluetooth SOC in 16.5mm^2 Silicon Area," Dig. Tech. Papers, Int. Solid-State Circuit Conf., pp. 90-91, Feb., 2002.

[10] F. Carrara, A. Scuderi, and G. Palmisano, "Wide-bandwidth Fully Integrated Cartesian Feedback Transmitter," Proc. IEEE Custom Integrated Circuits Conf., pp. 451-454, Sept., 2003.

[11] M.R. Elliott, T. Montalvo, B.P. Jeffries, F. Murden, J. Strange, A. Hill, S. Nandipaku, and J. Harrebek, "A Polar Modulator Transmitter for GSM/EDGE," IEEE J. Solid-State Circuits, vol 39, no. 12, pp. 2190-2199, Dec., 2004.

[12] P. Reynaert and M.S.J. Steyaert, "A 1.75GHz Polar Modulated CMOS RF Power Amplifier for GSM-EDGE," IEEE J. Solid-State Circuits, vol 40, no. 12, pp. 2598-2608, Dec., 2005.

[13] K.C. Tsai and P.R. Gray, “ A 1.9GHz, 1W CMOS Class-E Power Amplifier for Wireless Communications,” IEEE J. Solid-State Circuits, vol 34, no. 7, pp. 962-970, July, 1999.

[14] T. Sowlati and D.M.W. Leenaerts, “A 2.4-GHz 0.18-μm CMOS Self-Biased Cascode Power Amplifier,” IEEE J. Solid-State Circuits, vol 38, no. 8, pp. 1318-1324, Aug., 2003.

[15] M. Zargari, D.K. Su, C.P. Yue, S. Rabii, D. Weber, B.J. Kaczynski, S.S. Mehta, K. Singh, S. Mendis, and B.A. Wooley, “A 5-GHz CMOS Transceiver for IEEE 802.11a Wireless LAN Systems,” IEEE J. Solid-State Circuits, vol 37, no. 12, pp. 1688-1694, Dec., 2002.

[16] B. Razavi, RF Microelectronics, Upper Saddle River, NJ: Prentice-Hall, 1998.

[17] R.R. Gray and R.G. Meyer, Analysis and Design of Analog Integrated Circuits. New York: John Wiley, 1993.

[18] R.S. Narayanaswami, RF CMOS Class C Power Amplifiers for Wireless Communications, Ph.D. Dissertation, University of California, Berkeley, Electronics Research Laboratory, Memorandum No. UCB/ERL M01/73, 2001.

[19] L.R. Kahn, “Single-sideband Transmission by Envelope Elimination and Restoration,” Proc. IRE, pp. 803-806, July, 1952

[20] J.A. Weldon, 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, pp. 160-161, Feb., 2001.

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

[22] 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, pp. 155-160, May 1991.

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

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

[25] R. Gupta, and D.J. Allstot, “Fully Monolithic CMOS RF Power Amplifiers: Recent Advances,” IEEE Communications Magazine, vol. 37, issue 5, pp. 94-98, Spril, 1999.

[26] S.C. Cripps, “Advanced Techniques in RF Power Amplifier Design“, Norwood, MA: Artech House Incorporated, 2002.

[27] H. Chireix, “High Power Outphasing Modulation,” Proc. IRE, vol. 23, no.11, pp. 1370-1392, Nov.,1935.

[28] F.H. Raab, “Efficiency of Doherty RF Power Amplifier Systems,” IEEE Trans. on Broadcasting, vol. BC-33, no. 3, pp. 77-83, Sept. 1987.

[29] J. Vuolevi, and T. Rahkonen, “Distortion in RF Power Amplifiers,” Norwood, MA: Artech House Incorporated, 2003.

[30] F.H. Raab, et. Al, “Power Amplifiers and Transmitters for RF and Microwave,” IEEE Trans. on microwave theory and techniques, vol. 50, pp. 814-826, March, 2002.

[31] C. Fallesen, P. Asbeck, “A 1W CMOS Power Amplifier for GSM-1800 with 55% PAE,” Microwave Symposium Dig. Of Tech. papers, vol. 1, pp.453-456, May, 2001.

[32] I. Aoki, S.D. Kee, D.B. Rutledge, A. Hajimiri, “Fully Integrated CMOS Power Amplifier Design Using the Distribuited Active-Transformer Architecture,” IEEE Journal of Solid-State Circuits, vol. 37, pp. 371-383, March, 2002.

[33] I. Aoki, S. Kee, D. Rutledge, A. Hajimire, “A Fully-Integrated 1.8V, 2.8W, 1.9GHz CMOS Power Amplifier,” IEEE Radio Frequency Integrated Circuits Symposium, pp. 199-202, June, 2003.

[34] D. Su, W. McFarland, “A 2.5V, 1W monolithic CMOS RF Power Amplifier,” Proc. Of the IEEE Custom Integrated Curcuits Conference, pp. 189-192, May, 1997.

[35] J.C. Rudell, Frequency-Translations Techniques for High-Integration High-Selectivity Multi-standard Wireless Communication Systems, Ph.D. Thesis, University of California, Berkeley, 2000.

[36] 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. SC-10, pp. 168-176, June, 1975.

[37] F.H. Raab, “Idealized Operations of the Class E Tuned Power Amplifier,” IEEE Trans. on Circuits Systems, vol. CS-24, pp.725-735, 1977.

[38] D.M. Pozar, Microwave Engineering. Reading, MA: Addison-Wesley, 1993.