

UNIVERSITÀ DEGLI STUDI DI PAVIA
FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI ELETTRONICA



DOTTORATO DI RICERCA IN MICROELETTRONICA
XXII CICLO

EFFICIENCY ENHANCEMENT
TECHNIQUES FOR INTEGRATED POWER
AMPLIFIERS IN NEW GENERATION
CELLULAR APPLICATIONS

TUTORE:
PROF. DANILO MANSTRETTA

COORDINATORE:
CHIAR.MO PROF. RINALDO CASTELLO

TESI DI DOTTORATO DI
FLAVIO AVANZO

Index

| | |
|---|-----------|
| Index | ii |
| Introduction | v |
| Chapter 1..... | 1 |
| Systems for wireless applications..... | 1 |
| 1.1 Evolution of wireless mobile systems | 1 |
| 1.1.1 3G: UMTS | 2 |
| 1.1.2 3.5G: High Speed Packet Access (HSPA)..... | 10 |
| 1.1.3 4G: Long Term Evolution (LTE) | 11 |
| 1.2 Conclusion..... | 14 |
| Chapter 2..... | 17 |
| Linear Power Amplification and Efficiency Enhancement..... | 17 |
| 2.1 Linear Power Amplification | 17 |
| 2.2 Working Classes | 20 |
| 2.2.1 Class A..... | 21 |
| 2.2.2 Class B..... | 23 |
| 2.2.3 Class AB..... | 25 |
| 2.2.4 Harmonics generation..... | 25 |
| 2.3 Efficiency Enhancement Techniques | 26 |
| 2.3.1 Doherty Power Amplifier | 27 |
| 2.3.2 Envelope Tracking (Bias Adaptation) | 28 |
| 2.3.3 Chireix Amplifier (LINC) | 29 |
| 2.3.4 Envelope Elimination and Restoration | 31 |
| 2.4 Conclusion..... | 32 |
| Chapter 3..... | 35 |
| A Common Base Class AB PA Design and Testing..... | 35 |
| 3.1 Breakdown Mechanism in Bipolar Transistors | 35 |
| 3.2 Basic stages comparison..... | 38 |
| 3.2.1 Common Emitter | 38 |
| 3.2.2 Cascode..... | 40 |
| 3.2.3 Common Base | 41 |
| 3.2.4 AC-Coupled Cascode | 42 |
| 3.3 Class AB Common Base Design Example..... | 45 |
| 3.3.1 Common Base Output Stage | 45 |
| 3.3.2 Inter-stage Matching Network..... | 50 |
| 3.3.3 Common Emitter Driver Stage..... | 53 |
| 3.3.4 Bias Network | 55 |
| 3.3.5 Design Details | 57 |
| 3.3.6 Layout..... | 59 |
| 3.3.7 Test Board | 60 |
| 3.3.8 Simulation Results..... | 64 |
| 3.3.9 Measurement Setup | 65 |
| 3.3.10 Measurement Results..... | 66 |
| 3.4 Conclusion..... | 70 |

| | |
|---|------------|
| Chapter 4 | 73 |
| The Design of a CMOS Doherty Power Amplifier | 73 |
| 4.1 The Ideal Doherty Structure | 73 |
| 4.2 Second order effects | 77 |
| 4.2.1 Finite output resistance..... | 77 |
| 4.2.2 Phase mismatch | 79 |
| 4.2.3 Quarter-wave line | 80 |
| 4.3 Doherty Power Amplifier Design..... | 81 |
| 4.3.1 Main Amplifier Design and Impedance Transformer. | 81 |
| 4.3.2 Ideal Auxiliary Amplifier | 85 |
| 4.3.3 Class C Auxiliary Amplifier..... | 88 |
| 4.3.4 Output Impedance Transformation Network..... | 90 |
| 4.3.5 Pseudo Differential Solution | 94 |
| 4.3.6 Driver stage | 96 |
| 4.4 Conclusion..... | 98 |
| Thermal Effects | 101 |

Introduction

Terminals for mobile applications are nowadays of widespread use. In recent years we have assisted to a strong growth of the functionalities offered by these terminals, thanks also to the high bit rates offered by newly developed wireless communications standards, such as 3G cellular phones (UMTS) and WLANS. A stringent requirement for wireless terminals is the reduction of power consumption, aiming at a prolonged use of the mobile terminal. Contrary on the past, nowadays most of the energy stored in the battery is consumed during transmission. It follows that transmitter efficiency, in particular power amplifier efficiency, plays an important role in determining battery duration.

In order to allow for high bit rates for a given occupied bandwidth, modern communication standards use modulation schemes with high spectral efficiency. As a result, the modulated signal has a variable envelope and must be amplified by linear elements. The high peak-to-average power ratio typical of these signals forces the power amplifier (PA) to work several dBs below its peak output power, even when transmitting at maximum average output power. As a result, for a given average output power, the optimum load impedance and the amplifier average efficiency are significantly reduced compared to constant envelope operation. This problem is even more difficult to address using the most advanced silicon processes. In fact, the tremendous improvements in f_T and f_{MAX} of modern SiGe and SiGe:C technologies have been partly achieved exploiting the tradeoff between transit time and breakdown voltage: i.e. higher f_T and f_{MAX} have been exchanged for a reduction in breakdown voltage. The consequent reduction in supply voltage pushes the optimum load

impedance further down. As the load impedance level is reduced, the impedance transformation ratio from the PA output to the antenna 50Ω load increases, making the matching network more sensitive to components variations and parasitic elements that ultimately impact operative bandwidth and efficiency. To address these issues, a common-base topology that is able to sustain output voltages well in excess of the collector-emitter breakdown voltage (BV_{CEO}) has been developed in the first part of this research work. A class AB power amplifier based upon this topology has been designed and tested.

As already anticipated the main disadvantage in using linear power amplifiers is that their efficiency drops dramatically if the power of the signal is decreased from the peak value. Since the variable envelope signals have a mean power much lower than the peak power, average efficiency of the amplifier is rather low. This limitations can be overcome using efficiency enhancement techniques. One of these is the Doherty amplifier, developed in the early 30's of the past century at the Bell Laboratories. It is based upon a pair of power amplifiers (main and auxiliary) coupled to each other with an impedance inverter network. This technique allows a linear power amplifier to reach the maximum efficiency over a wider range of output power if compared to a classic linear amplifier. The auxiliary amplifier turns on only when the output power exceeds a certain amount and allows to reduce the load impedance seen by the main amplifier. This allows the main amplifier to work under maximum efficiency conditions with high linearity. Moreover this efficiency enhancement technique is intrinsically wideband, on the contrary of other efficiency techniques that are band-limited. The second part of this thesis will deal with the design of a fully integrated Doherty Power Amplifier able to deliver a maximum linear power of 30dBm.

In the *first chapter* the 3G (and beyond) communication standards are introduced, posing the attention to the capabilities of the modulation systems and the requirements of the power amplifier.

In the *second chapter* the working principle of a linear power amplifier and its performance in terms of efficiency is compared among different linear amplification classes. Also, the problem of efficiency enhancement will be addressed showing several techniques able to address this problem.

In the *third chapter* the design and testing of a class AB linear power amplifier in a Si:Ge 0.25 μ m technology in a common base topology is considered. The performance comparison among different amplification topologies is discussed, posing the attention to the benefits introduced by a passive current amplification.

The *fourth chapter* deals with the design of a Doherty PA, posing the attention to the Auxiliary PA, which has the major role in the structure performance. A solution which allows to employ a modulator which conveniently varies the phase shift at the signal apply to the main and the peaking amplifier is be considered, allowing to integrate the overall transmitter.

Chapter 1

Systems for wireless applications

This chapter will cover the evolution of the wireless transmission systems over the various technological eras. An overview of third generation cellular systems and later generations will be shown.

1.1 Evolution of wireless mobile systems

In the last decades an exponential growth in the number of users of mobile phones took place, moving the research to overcome the limits related to each mobile generation.

The first generation mobile networks were based on a pure analog modulation. In those systems the territory was divided in several areas (named cells) where the communication took place, even if the user was moving inside it. The frequency band was divided into several channels, where the total channel bandwidth was available for the user. Thus this system suffered of a low capability to support a large number of users. Moreover mobility between different countries was limited, because the service wasn't supplied beyond the country where the contract was issued. Due to the analog modulation used, security issues were present because of the absence of cryptographic encoding.

In the second generation systems (still in use today) a digital modulation scheme is employed. This allows a higher security during the data/voice link thanks to the coded information. The GSM (Global System for Mobile communication) was the first mobile worldwide network and it is still available in two versions (GSM900 and GSM1800). It supplies 992 channels (voice and signalling) with a Frequency Division Duplex (890÷915 MHz uplink and 935÷960 MHz downlink). This system allows a TDM and FDM access and the modulation used is the GMSK. Data communication is also available, with a maximum data rate of 9.6 kbit/s.

The need for faster data rates has facilitated the development of an intermediate generation between the second and the third one: the 2.5 generation. This generation is based on the well-established GSM technology. Thanks to the impulse due to the growth of the Internet web, this generation allows a faster data rate service over a radio link. The GPRS standard was then born by a slight modification of the GSM network already present, supplying a packet switching network suitable for data exchange. The GPRS system allows a data rate of 171.2 kbit/s.

Also the EDGE (Enhanced Data rate for GSM Evolution) system can be included in this intermediate generation: this radio link technology has a more efficient use of the available GSM FDMA channel bandwidth using a multitasking modulation: respect to the GMSK it is possible to triple the bit rate for the same occupied bandwidth. Thus EDGE is able to reach a maximum bit rate of at least 384kbit/s for a user which travels at a speed until 100km/h and 144kbit/s for a 250km/h.

The third generation of mobile devices is now growing, having the advantages of high flexibility, very high data rates and integration with the fixed network. The UMTS (Universal Mobile Telecommunications System) standard employs a combination of TDMA, FDMA and CDMA with a direct-sequence spread-spectrum (DS-SS) digital modulation. The main difference between this system and the previous generations refers to the employment of a new radio interface which allows higher maximum data rates (up to 2Mbit/s for the W-CDMA interface) but offering also intermediate rates. The gradual transition to this generation is possible thanks to the compatibility with the previous standards.

1.1.1 3G: UMTS

The third generation standard for cellular applications nowadays used is the UMTS. This is the necessary evolution of the previous GSM system, and its presence into the market took place several years ago. Differently from GSM (which was intended essentially for voice applications) the UMTS is a data-oriented standard, allowing a fast connection to the Web and its services like video-calls, e-mailing, data base searching and streaming. Another difference between this standard and the second generation is the type of access. While GSM was based only to TDMA and FDMA, UMTS is based also on a Code Division Multiple Access (CDMA). Let's look more in detail how the different access modalities work.

Three main resources are available during a wireless communication: frequency, time and power. Different users can be distinguished to each other if a different portion of the available resource is assigned to them. This three resources can be depicted (Figure 1.1) as a three dimensional space, where the user can be described by three references (frequency, time and power).

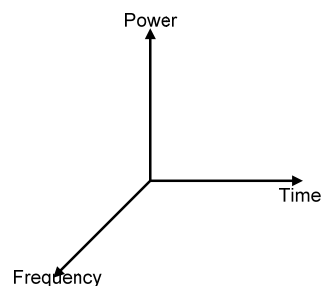


Figure 1.1 *Three dimensional space of the available resources for transmission*

In the TDMA access the shared resource is the time. Thus, in three dimensional space above described, each user is located into a region where power and frequency are completely available, but just for a fixed amount of time (Figure 1.2).

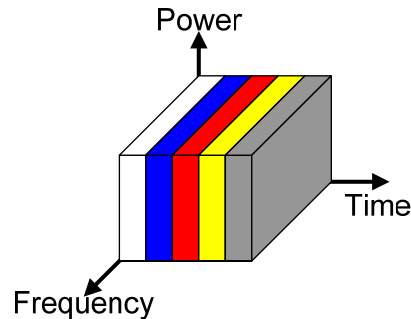


Figure 1.2 *TDMA access*

In the simplest version the transmission period is divided into time frames of the same length, each of them is divided into a fixed number of time slots. The time slots are assigned to different users, and this allocation is fixed for all the frames. This means that a single user can transmit only during the assigned time slot: here the total channel bandwidth and base station power are available. This resources assignment has a drawback of a non efficient channel use. This is because each user has the same quantity of resources available (one time slot for frame) regardless the quantity of data it has to transmit. Thus the quantity of data available in a time slot must be sufficient for the user that generates the most of the traffic; while the other users (which generate less traffic) have a larger amount of available resource even if they generate less traffic. This is a waste of channel capability.

In the FDMA the shared resource is the frequency. This means that each user has a portion of the channel bandwidth available for all the time with all of the available power (Figure 1.3)

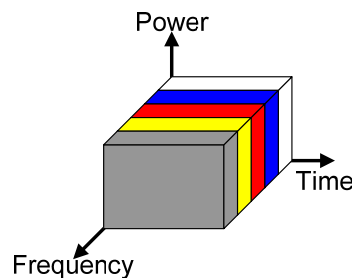


Figure 1.3 *FDMA access*

The FDMA protocol divides the overall channel bandwidth into a certain amount of frequency band, each of them dedicated to a single user. The FDMA has the same drawback like the TDMA in terms of low channel capability: if a user doesn't need to transmit for a certain amount of time, its bandwidth cannot be shared by another user. Moreover guard band are present between different

channels, increasing the wasted bandwidth. Compared to the TDMA this is a less efficient access, but it is simpler due to the absence of synchronization.

The use of both TDMA and FDMA avoids interferences between the two domains, because they are orthogonal to each other. A third multiplication method allows complete superposition of TDMA and FDMA exploiting the orthogonality in the power domain. This multiplication method is depicted in (Figure 1.4) where is evident that the total power of the base station is divided among different users.

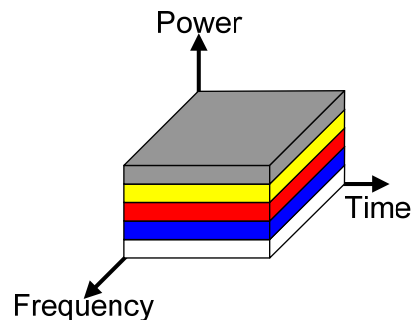


Figure 1.4 *CDMA access*

The process of signal encoding (assigning different codes for different users) affects the signal spectrum: since the code bandwidth is larger than the signal bandwidth the encoding process spreads the signal spectrum. This is the reason why this operation is called *spread spectrum*, and the protocols based upon CDMA are called Spread Spectrum Multiple Access (SSMA). During transmission the signal spectrum is spread over a wider frequency band, thus reducing spectral density. This spectrum will be then concentrated again during reception.

Spread spectrum signals are obtained by modulating the data signal with a unique code which is assigned to each. The encoding procedure can be realized in different ways. Depending on the modulation technique used it is possible to identify several encoding methods:

- *Direct Sequence Spread Spectrum*: the high speed spreading code is directly multiplied with the data signal.
- *Frequency Hopping Spread Spectrum*: the frequency carrier at which the data signal is modulated is rapidly changed due to the spreading code.
- *Time Hopping Spread Spectrum*: the data signal is not continuously transmitted but it is split in different bursts whose temporal position is set by the spreading code.

The UMTS uses the first method and the resulting transmission is further multiplied with a scrambling code which increases the orthogonality among users without affecting the data rate. This spreading and scrambling procedure is shown in Figure 1.5.

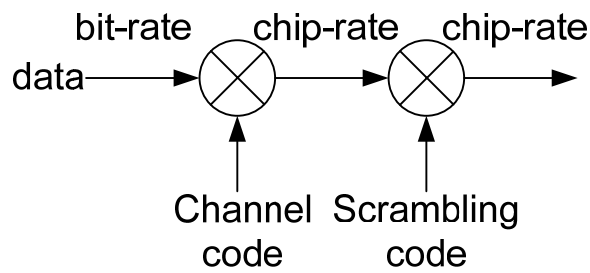


Figure 1.5 Direct Sequence (DS) spread spectrum

Looking only at the spreading procedure, it is possible to depict how the data signal is modified in the time and frequency domain (Figure 1.6 and Figure 1.7)

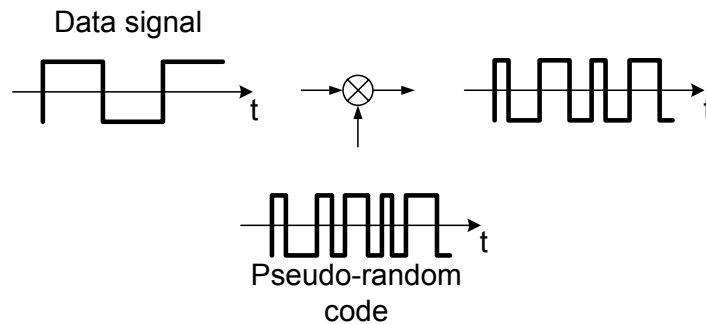


Figure 1.6 DS spread spectrum in the time domain

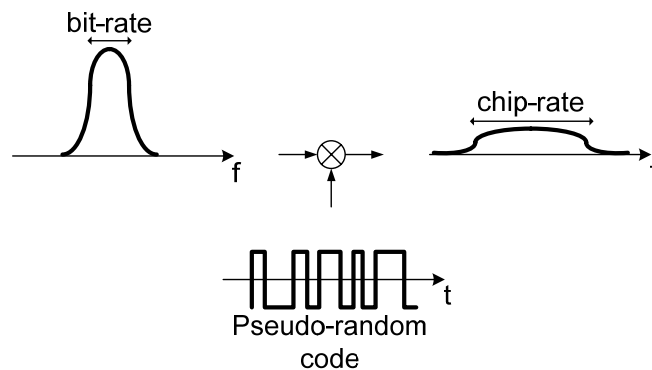


Figure 1.7 DS spread spectrum in the frequency domain

The modulation used is the dual-code BPSK, which modulation scheme is drawn in Figure 1.8. The DPDCH is a data channel, while DPCCH is a control channel. In this modulation the data and control channels are then transmitted on the in-phase (I) and quadrature channel (Q) respectively. The spreading factor indicates the number of orthogonal channels. The spreading factor is 256 for the control channel, while it is variable for the data channel (depending on the desired data rate).

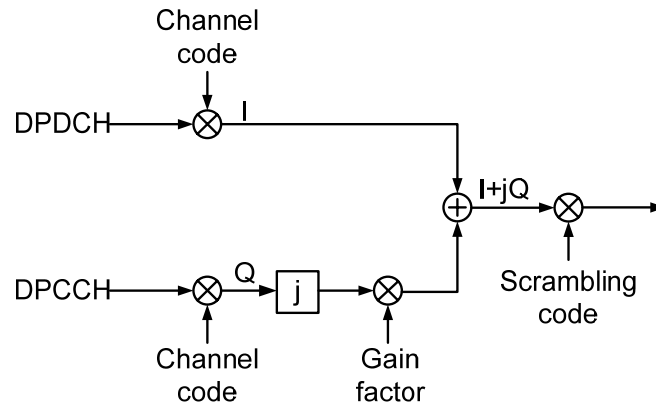


Figure 1.8 Dual-code BPSK modulation scheme

The trajectory of the UMTS modulated signal over the I-Q plane and its spectrum are shown in Figure 1.9.

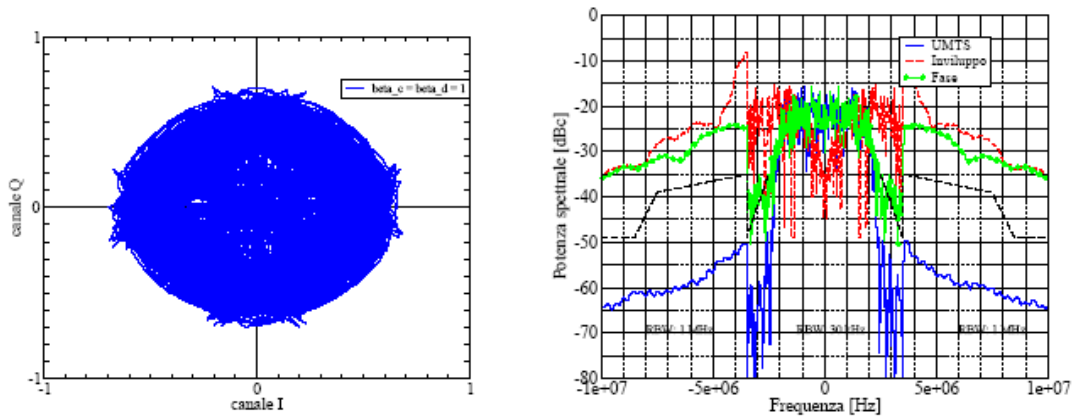


Figure 1.9 Trajectory and spectrum of the UMT signal

The signal can have null amplitude (since the trajectory crosses the axes origin) and thus null signal power. The modulated signal has a non constant envelope thus a linear amplification of this signal is necessary during transmission. It has an average power lower than the peak output power. The *Crest Factor (CF)* (also named Peak to Average Power Ratio – PAPR) defines the ratio between the peak and the rms value of the modulated signal envelope power:

$$[1.1] \quad CF = \frac{|E_{\max}|^2}{RMS}$$

where

$$[1.2] \quad RMS = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T |E(t)|^2 dt$$

The CF for a W-CDMA signal is variable, depending on the number of code channels, with a maximum value of 4.5dB (Table 1.1) [1]. Looking at the probability density function for a UMTS signal (Figure 1.10) it is possible to see that it is maximum when the envelope has an amplitude of 0.7 times of the peak value. This again shows PA works most of the time well below the peak output power, requiring a linear amplification with an efficiency maximized at the average power. Thus the power amplifier used should have an high efficiency over a wide output power range and not just at the maximum transmittable power.

| N. of Code Channels | PAPR |
|---------------------|-------|
| 1 | 4.5dB |
| 4 | 9dB |
| 16 | 10dB |
| 128 | 11dB |

Table 1.1 PAPR for a different amount of code channels

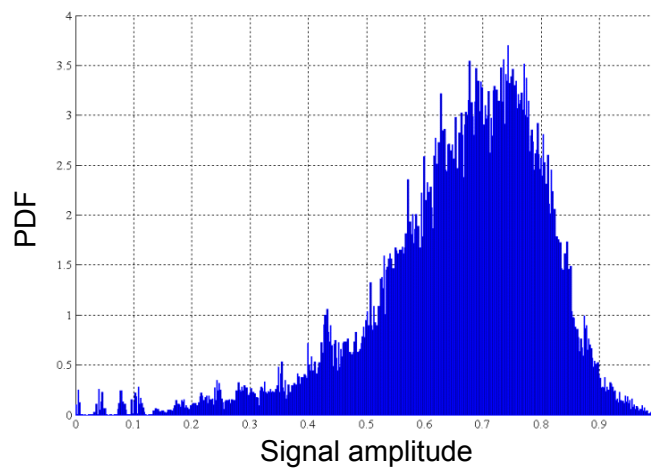


Figure 1.10 Probability Density function for an UMTS modulated signal

The characteristics which have to be satisfied by a transmitter for UMTS applications are reported below [2].

Frequency Bands. The frequency bands allowed for the UTRA/FDD are reported in Table 1.2.

| Operating Band | UL Frequencies | DL frequencies |
|----------------|-----------------------------|-----------------------------|
| | UE transmit, Node B receive | UE receive, Node B transmit |
| I | 1920 - 1980 MHz | 2110 -2170 MHz |
| II | 1850 -1910 MHz | 1930 -1990 MHz |

Table 1.2 Frequency bands allowed for the UTRA/FDD

Occupied Bandwidth. Occupied bandwidth is a measure of the bandwidth containing 99% of the total integrated power of the transmitted spectrum, centered on the assigned channel frequency. The occupied channel bandwidth shall be less than 5 MHz based on a chip rate of 3.84 Mcps.

UE maximum output Power. The following Power Classes (Table 1.3) define the nominal maximum output power. The nominal power is the broadband transmit power of the UE, i.e. the power in a bandwidth of at least $(1+\alpha)$ times the chip rate of the radio access mode. The period of measurement shall be at least one timeslot.

Out of Band Emissions. Out of band emissions are unwanted emissions immediately outside the nominal channel resulting from the modulation process and non-linearity in the transmitter but excluding spurious emissions. This out of band emission limit is specified in terms of a spectrum emission mask and Adjacent Channel Leakage power Ratio.

| Power Class | Nominal maximum output power | Tolerance |
|-------------|------------------------------|------------|
| 1 | +33 dBm | +1/-3 dB |
| 2 | +27 dBm | +1/-3 dB |
| 3 | +24 dBm | +1/-3 dB |
| 4 | +21 dBm | ± 2 dB |

Table 1.3 UMTS power requirements

Spectrum Emission Mask. The spectrum emission mask (Table 1.4) of the UE applies to frequencies, which are between 2.5 MHz and 12.5 MHz away from the UE centre carrier frequency. The out of channel emission is specified relative to the RRC filtered mean power of the UE carrier.

| Δf^* in MHz | Minimum requirement | Additional Minimum requirement for operation in Band b | Measurement bandwidth |
|--|---|--|-----------------------|
| 2.5 - 3.5 | $\left\{ -35 - 15 \cdot \left(\frac{\Delta f}{\text{MHz}} - 2.5 \right) \right\} \text{dBc}$ | -15 dBm | 30 kHz ** |
| 3.5 - 7.5 | $\left\{ -35 - 1 \cdot \left(\frac{\Delta f}{\text{MHz}} - 3.5 \right) \right\} \text{dBc}$ | -13 dBm | 1 MHz *** |
| 7.5 - 8.5 | $\left\{ -39 - 10 \cdot \left(\frac{\Delta f}{\text{MHz}} - 7.5 \right) \right\} \text{dBc}$ | -13 dBm | 1 MHz *** |
| 8.5 - 12.5 MHz | -49 dBc | -13 dBm | 1 MHz *** |
| * Δf is the separation between the carrier frequency and the centre of the measuring filter. | | | |
| ** The first and last measurement position with a 30 kHz filter is at Δf equals to 2.515 MHz and 3.485 MHz. | | | |
| *** The first and last measurement position with a 1 MHz filter is at Δf equals to 4 MHz and 12 MHz. As a general rule, the resolution bandwidth of the measuring equipment should be equal to the measurement bandwidth. To improve measurement accuracy, sensitivity and efficiency, the resolution bandwidth can be different from the measurement bandwidth. When the resolution bandwidth is smaller than the measurement bandwidth, the result should be integrated over the measurement bandwidth in order to obtain the equivalent noise bandwidth of the measurement bandwidth. | | | |
| The lower limit shall be -50 dBm/3.84 MHz or which ever is higher. | | | |

Table 1.4 UMTS spectrum emission mask

Adjacent Channel Leakage power Ratio (ACLR). This is the ratio of the RRC filtered mean power centered on the assigned channel frequency to the RRC filtered mean power centered on an adjacent channel frequency. If the adjacent channel power is greater than -50dBm then the ACLR shall be higher than the value specified in Table 1.5.

| Power Class | Adjacent channel frequency relative to assigned channel frequency | ACLR limit |
|-------------|---|------------|
| 3 | + 5 MHz or - 5 MHz | 33 dB |
| 3 | + 10 MHz or - 10 MHz | 43 dB |
| 4 | + 5 MHz or - 5 MHz | 33 dB |
| 4 | + 10 MHz or -10 MHz | 43 dB |

Table 1.5 ACLR requirements

Spurious Emissions. Spurious emissions are emissions which are caused by unwanted transmitter effects such as harmonics emission, parasitic emission, intermodulation products and frequency conversion products, but exclude out of band emissions. The minimum requirements (Table 1.6) are only applicable for frequencies, which are greater than 12.5 MHz away from the UE centre carrier frequency. Additional requirements are shown in Table 1.7.

| Frequency Bandwidth | Measurement Bandwidth | Minimum requirement |
|--|-----------------------|---------------------|
| $9\text{ kHz} \leq f < 150\text{ kHz}$ | 1 kHz | -36 dBm |
| $150\text{ kHz} \leq f < 30\text{ MHz}$ | 10 kHz | -36 dBm |
| $30\text{ MHz} \leq f < 1000\text{ MHz}$ | 100 kHz | -36 dBm |
| $1\text{ GHz} \leq f < 12.75\text{ GHz}$ | 1 MHz | -30 dBm |

Table 1.6 Spurious emissions requirements

| Paired band | Frequency Bandwidth | Measurement Bandwidth | Minimum requirement |
|--|---|-----------------------|---------------------|
| For operation in frequency bands as defined in subclause 5.2(a) | $1884.5\text{ MHz} < f < 1919.6\text{ MHz}$ | 300 kHz | -41 dBm |
| | $925\text{ MHz} \leq f \leq 935\text{ MHz}$ | 100 kHz | -67 dBm * |
| | $935\text{ MHz} < f \leq 960\text{ MHz}$ | 100 kHz | -79 dBm * |
| | $1805\text{ MHz} \leq f \leq 1880\text{ MHz}$ | 100 kHz | -71 dBm * |
| NOTE *: The measurements are made on frequencies which are integer multiples of 200 kHz. As exceptions, up to five measurements with a level up to the applicable requirements defined in Table 6.12 are permitted for each UARFCN used in the measurement | | | |

Table 1.7 Additional requirements for spurious emissions

Error Vector Magnitude. The Error Vector Magnitude is a measure of the difference between the reference waveform and the measured waveform. This difference is called the error vector. Both waveforms pass through a matched Root Raised Cosine filter with bandwidth 3,84 MHz and roll-off $\alpha=0,22$. Both waveforms are then further modified by selecting the frequency, absolute phase, absolute amplitude and chip clock timing so as to minimize the error vector. The EVM result is defined as the square root of the ratio of the mean error vector power to the mean reference power expressed as a %. The Error Vector Magnitude shall not exceed 17.5 % for the parameters specified in Table 1.8.

| Parameter | Unit | Level |
|-------------------------|------|-------------------|
| UE Output Power | dBm | ≥ -20 |
| Operating conditions | | Normal conditions |
| Power control step size | dB | 1 |

Table 1.8 EVM requirements

1.1.2 3.5G: High Speed Packet Access (HSPA)

The increasing demand for higher data capability for the wireless communication standards, has led to the development of new protocols able to extend the performance of the existing WCDMA protocols. The first UMTS Release 99 (R99) [2] allowed a 384 kbit/s downlink and uplink maximum throughput, which quickly became insufficient to cover the growing demand of new services. HSPA refers to an improved version of the early UMTS and concerns the High Speed Downlink Packet Access (HSDPA) and the High Speed Uplink Packet Access (HSUPA). These two protocols provide increased performance by using improved modulation schemes and reducing the latency. HSPA is able to offer a 14.4 Mbps in downlink and up to 5.5 Mbps in uplink. Both HSDPA and HSUPA can be implemented in the standard 5MHz carrier of the UMTS networks and co-exist with the first generation of UMTS networks based on the 3GPP R99 standards.

The improved speed and latency of HSPA offers a much improved end-user experience for services such as Multimedia Messaging, Mobile TV, e-mail and music downloads. Figure 1.11 shows the increasing services available with this improved standard and those available with the next generation of mobile standards (4G LTE) which will be introduced later.

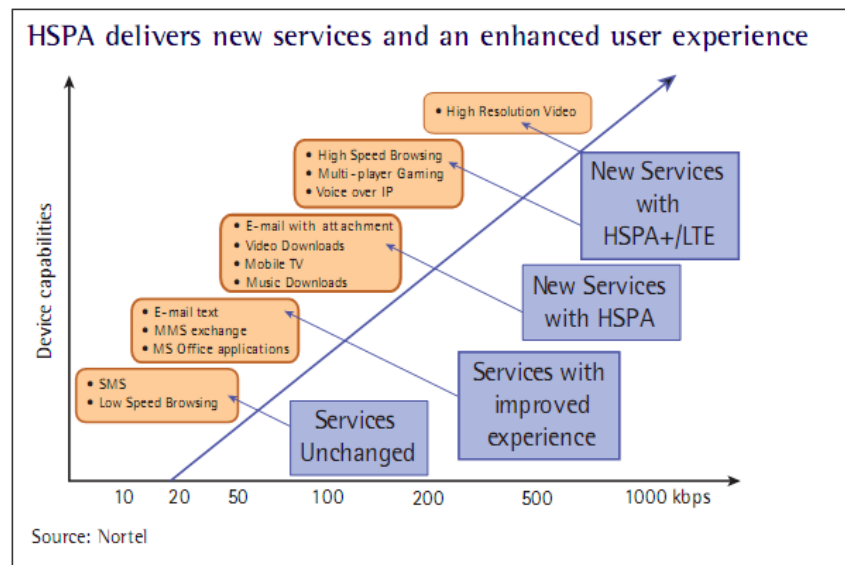


Figure 1.11 Service available for the next generation mobile standards

As HSPA standards refer uniquely to the access network, there is no change required of the core network: thus HSPA just requires new terminals in order to be able to co-exist with the older standards. HSUPA defines a new radio interface for the uplink communication. The overall goal is to improve the coverage and throughput as well as to reduce the delay of the uplink dedicated transport channels. From a 3GPP point of view, the first set of standards was approved in December 2004, and performance aspects were finalized during the summer of 2005.

HSUPA uses an uplink enhanced dedicated channel (E-DCH). Since this is just a software update for the handset, the transmitter characteristics are unchanged compared to the previous UMTS release.

1.1.3 4G: Long Term Evolution (LTE)

The LTE is the last step toward the 4th generation (4G) of radio technologies designed to increase the capacity and speed of mobile telephone networks. Where the current generation of mobile telecommunication networks are collectively known as 3G (for third generation), LTE is marketed as 4G. LTE is a set of enhancements to the UMTS which will be introduced in 3GPP release 8.

The LTE is a wireless broadband internet system. It is an all-IP network based upon TCP/IP, with higher level services such as voice, video and messaging. It is significantly different if compared to the UMTS/HSPA, since it is designed principally for data communications instead of voice communications. LTE is backward compatible with non-3GPP as well as 3GPP technologies. Its ability to interwork with legacy and new networks, and its seamless integration of Internet applications will drive the convergence between fixed and mobile systems (Figure 1.12) and facilitates new type of services (Figure 1.11). LTE heralds a new era with the transition from circuit switched approaches for voice traffic to a fully packet switched model.

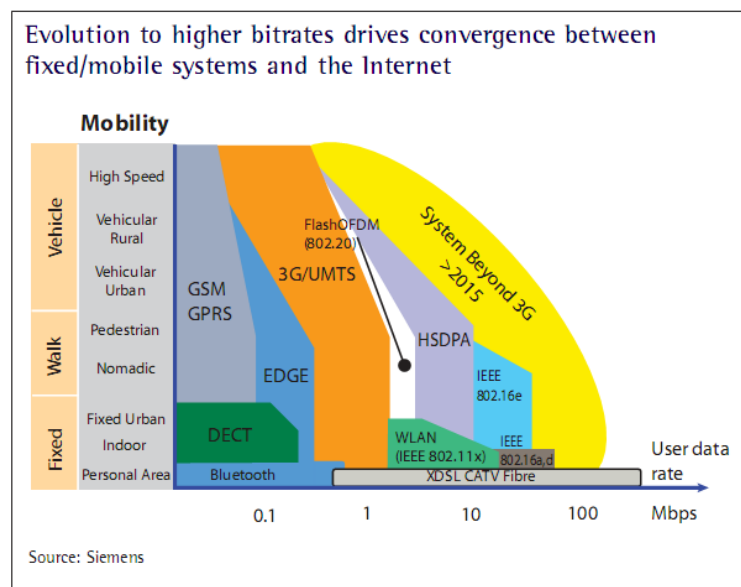


Figure 1.12 Convergence between fixed and mobile systems

LTE contains a new radio interface and access network designed to deliver higher data rates (up to peak rates of 75 Mbps on the uplink and 300 Mbps on the downlink) and fast communication times. The technology chosen by 3GPP for the LTE air interface uses Orthogonal Frequency Division Multiplexing (OFDM) and MIMO technologies, together with high data rate modulation.

OFDM-based technologies can achieve the targeted high data rates with simpler implementations involving relatively low cost and power-efficient hardware.

Data rates in WCDMA networks are constrained by the 5 MHz channel width. LTE overcomes these limitations by deploying in bandwidths up to 20 MHz. At bandwidths below 10 MHz, HSPA and LTE provide similar performance for the same number of antennas. Use of a wider RF band such as 20 MHz leads to group delay problems in WCDMA that limit the achievable data rate. LTE removes these limitations by deploying OFDM technology to split the 20 MHz channel into many narrow sub-channels. Each narrow sub-channel is driven to its maximum and the sub-channels subsequently combined to generate the total data throughput. LTE uses OFDMA in the downlink but Single Carrier-Frequency Division Multiple Access (SC-FDMA) in the uplink. SC-FDMA is technically similar to OFDMA but is better suited for handheld devices because it is less demanding on battery power. In fact it has a lower Peak to Average Power Ratio compared to the standard OFDM modulation, allowing to use linear power amplifier closer to their maximum efficiency performance.

As in OFDMA, the transmitters in an SC-FDMA system use different orthogonal frequencies (subcarriers) to transmit information symbols. However, they transmit the subcarriers sequentially, rather than in parallel. Relative to OFDMA, this arrangement reduces considerably the envelope fluctuations in the transmitted waveform. The throughput depends on the way in which information symbols are applied to subcarriers. There are two approaches to apportioning subcarriers among terminals.

In localized SC-FDMA (LFDMA), each terminal uses a set of adjacent subcarriers to transmit its symbols. Thus the bandwidth of an LFDMA transmission is confined to a fraction of the system bandwidth. The alternative to LFDMA is distributed SC-FDMA in which the subcarriers used by a terminal are spread over the entire signal band. One realization of distributed SC-FDMA is interleaved FDMA (IFDMA) where occupied subcarriers are equidistant from each other. A summary of the PAPR for LFDMA e IFDMA compared to the standard OFDMA including some pulse shaping with a RRC pulse shaping is shown in Table 1.9. The SC-FDMA shows lower PAPR compared to the OFDMA, and its PAPR is comparable with that of the W-CDMA (Table 1.1) [5].

| Modulation format | LFDMA | | | IFDMA | | | OFDMA |
|-------------------|------------------|-----------------------------|------------------------------|------------------|-----------------------------|------------------------------|--------|
| | No pulse shaping | Pulse shaping (rolloff 0.5) | Pulse shaping (rolloff 0.22) | No pulse shaping | Pulse shaping (rolloff 0.5) | Pulse shaping (rolloff 0.22) | |
| QPSK | 7.5dB | 7.6dB | 7.6dB | 0dB | 4.3dB | 6.1dB | 10.7dB |
| 8QPSK | 7.4dB | 7.5dB | 7.5dB | 0dB | 4.2dB | 5.9dB | 10.6dB |
| 16QAM | 8.4dB | 8.4dB | 8.5dB | 3.5dB | 6.6dB | 7.7dB | 10.5dB |
| 32QAM | 8.2dB | 8.3dB | 8.4dB | 3.4dB | 6.4dB | 7.5dB | 10.6dB |
| 64QAM | 8.6dB | 8.7dB | 8.7dB | 4.8dB | 7.1dB | 8.0dB | 10.5dB |

Table 1.9 PAPR of the OFDMA modulation

The frequency bands for the LTE standard are reported in Table 1.10, while the maximum power requirements for the power amplifier are reported in Table 1.11. It has to be noted that the maximum power levels are lower compared to the UMTS requirements (Table 1.3). Since the PAPR can be up to 8-9dB, the maximum power that the power amplifier must deliver is around 31-32dBm ± 2 dB.

| E-UTRA Operating Band | Downlink | | | Uplink | | |
|-----------------------------|---------------------------|----------------------|--------------------------|---------------------------|----------------------|--------------------------|
| | F _{DL_low} (MHz) | N _{Offs-DL} | Range of N _{DL} | F _{UL_low} (MHz) | N _{Offs-UL} | Range of N _{UL} |
| 1 | 2110 | 0 | 0 – 599 | 1920 | 18000 | 18000 – 18599 |
| 2 | 1930 | 600 | 600 – 1199 | 1850 | 18600 | 18600 – 19199 |
| 3 | 1805 | 1200 | 1200 – 1949 | 1710 | 19200 | 19200 – 19949 |
| 4 | 2110 | 1950 | 1950 – 2399 | 1710 | 19950 | 19950 – 20399 |
| 5 | 869 | 2400 | 2400 – 2649 | 824 | 20400 | 20400 – 20649 |
| 6 | 875 | 2650 | 2650 – 2749 | 830 | 20650 | 20650 – 20749 |
| 7 | 2620 | 2750 | 2750 – 3449 | 2500 | 20750 | 20750 – 21449 |
| 8 | 925 | 3450 | 3450 – 3799 | 880 | 21450 | 21450 – 21799 |
| 9 | 1844.9 | 3800 | 3800 – 4149 | 1749.9 | 21800 | 21800 – 22149 |
| 10 | 2110 | 4150 | 4150 – 4749 | 1710 | 22150 | 22150 – 22749 |
| 11 | 1475.9 | 4750 | 4750 – 4999 | 1427.9 | 22750 | 22750 – 22999 |
| 12 | 728 | 5000 | 5000 – 5179 | 698 | 23000 | 23000 – 23179 |
| 13 | 746 | 5180 | 5180 – 5279 | 777 | 23180 | 23180 – 23279 |
| 14 | 758 | 5280 | 5280 – 5379 | 788 | 23280 | 23280 – 23379 |
| ... | | | | | | |
| 17 | 734 | 5730 | 5730 – 5849 | 704 | 23730 | 23730 – 23849 |
| ... | | | | | | |
| 33 | 1900 | 36000 | 36000 – 36199 | 1900 | 36000 | 36000 – 36199 |
| 34 | 2010 | 36200 | 36200 – 36349 | 2010 | 36200 | 36200 – 36349 |
| 35 | 1850 | 36350 | 36350 – 36949 | 1850 | 36350 | 36350 – 36949 |
| 36 | 1930 | 36950 | 36950 – 37549 | 1930 | 36950 | 36950 – 37549 |
| 37 | 1910 | 37550 | 37550 – 37749 | 1910 | 37550 | 37550 – 37749 |
| 38 | 2570 | 37750 | 37750 – 38249 | 2570 | 37750 | 37750 – 38249 |
| 39 | 1880 | 38250 | 38250-38649 | 1880 | 38250 | 38250-38649 |
| 40 | 2300 | 38650 | 38650-39649 | 2300 | 38650 | 38650-39649 |

NOTE: The channel numbers that designate carrier frequencies so close to the operating band edges that the carrier extends beyond the operating band edge shall not be used. This implies that the first 7, 15, 25, 50, 75 and 100 channel numbers at the lower operating band edge and the last 6, 14, 24, 49, 74 and 99 channel numbers at the upper operating band edge shall not be used for channel bandwidths of 1.4, 3, 5, 10, 15 and 20 MHz respectively.

Table 1.10 Frequency bands allowed for the LTE

| EUTRA band | Class 1 (dBm) | Tolerance (dB) | Class 2 (dBm) | Tolerance (dB) | Class 3 (dBm) | Tolerance (dB) | Class 4 (dBm) | Tolerance (dB) |
|------------|---|----------------|---------------|----------------|---------------|-----------------|---------------|----------------|
| 1 | | | | | 23 | ±2 | | |
| 2 | | | | | 23 | ±2 ² | | |
| 3 | | | | | 23 | ±2 ² | | |
| 4 | | | | | 23 | ±2 | | |
| 5 | | | | | 23 | ±2 | | |
| 6 | | | | | 23 | ±2 | | |
| 7 | | | | | 23 | ±2 ² | | |
| 8 | | | | | 23 | ±2 ² | | |
| 9 | | | | | 23 | ±2 | | |
| 10 | | | | | 23 | ±2 | | |
| 11 | | | | | 23 | ±2 ² | | |
| 12 | | | | | 23 | ±2 ² | | |
| 13 | | | | | 23 | ±2 | | |
| 14 | | | | | 23 | ±2 | | |
| 17 | | | | | 23 | ±2 | | |
| ... | | | | | | | | |
| 33 | | | | | 23 | ±2 | | |
| 34 | | | | | 23 | ±2 | | |
| 35 | | | | | 23 | ±2 | | |
| 36 | | | | | 23 | ±2 | | |
| 37 | | | | | 23 | ±2 | | |
| 38 | | | | | 23 | ±2 | | |
| 39 | | | | | 23 | ±2 | | |
| 40 | | | | | 23 | ±2 | | |
| Note 1: | The above tolerances are applicable for UE(s) that support up to 4 E-UTRA operating bands. For UE(s) that support 5 or more E-UTRA bands the maximum output power is expected to decrease with each additional band and is FFS | | | | | | | |
| Note 2: | For transmission bandwidths (Figure 5.6-1) confined within F_{UL_low} and $F_{UL_low} + 4$ MHz or $F_{UL_high} - 4$ MHz and F_{UL_high} , the maximum output power requirement is relaxed by reducing the lower tolerance limit by 1.5 dB | | | | | | | |

Table 1.11 Power requirements for the LTE

1.2 Conclusion

In this chapter the main characteristics of the 3G and beyond cellular standards have been introduced. These systems employ a modulation scheme which generate a signal with a variable envelope. Thus the average output power that must be transmitted is well below the maximum power, and it is determined by the Peak to Average Power Ratio. The power amplifier used to transmit these signals must be able to deliver an average output power as stated by the standard. Moreover the amplification must be performed with a good linearity, in order to fulfill the requirements of EVM and ACPR

References

- [1] Elisabetta De Bernardi di Valserra, “Architetture e circuiti per ricevitori integrati per lo standard UMTS in tecnologia CMOS”, *Pavia, Tesi di laurea, A.A. 1999/2000*
- [2] 3GPP TS 25.101 V3.19.0 (2006-12) Technical Specification Group Radio Access Networks; User Equipment (UE) radio transmission and reception (FDD) (Release 1999)
- [3] 3GPP TS 25.101 V6.19.0 (2009-03) Technical Specification Group Radio Access Network; User Equipment (UE) radio transmission and reception (FDD) (Release 6)
- [4] Pretl, H.; Maurer, L.; Schelmbauer, W.; Weigel, R.; Adler, B.; Fenk, J.; “Linearity considerations of W-CDMA front-ends for UMTS”, *Microwave Symposium Digest., 2000 IEEE MTT-S International Volume 1, 11-16 June 2000 Page(s):433 - 436 vol.1.*
- [5] Myung, H.G.; Junsung Lim; Goodman, D.J.; “Peak-To-Average Power Ratio of Single Carrier FDMA Signals with Pulse Shaping”, *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on 11-14 Sept. 2006 Page(s):1 - 5*

Chapter 2

Linear Power Amplification and Efficiency Enhancement

The power amplifier is the element which consumes the largest amount of power compared to the rest of the transmitter in an handset. The power available from the batteries is used to amplify the input signal that has to be transmitted. The operation which converts the DC power supplied by the batteries of the handset into RF power has to be performed in the most efficient way. Obviously this power cannot be completely transformed into a signal power: the remaining part is converted in heat. This chapter will describe the working principle of a power amplifier and the performance in terms of efficiency will be compared among different linear amplification classes. Moreover, the problem of efficiency enhancement will be addressed, showing several techniques which are able to address this problem.

2.1 Linear Power Amplification

As shown in the previous chapter, the 3G (and later) wireless communication standards employ a modulation system with high spectral efficiency in order to increase the data rate. This modulation generates a signal with a non constant envelope, which needs to be linearly amplified in order to preserve the information in the amplitude variation. Thus, in order to transmit this signal a linear power amplifier is needed.

A power amplifier is named “linear” when it linearly amplifies the input signal power. This doesn’t mean that signal harmonics are not generated by the amplifier. A proper resonant matching network resonates out the harmonics in order to prevent them flowing through the antenna. The simplified scheme of a prototype linear power amplifier is shown in Figure 2.1.

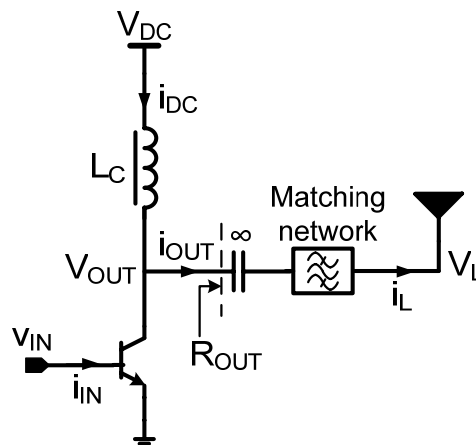


Figure 2.1 Prototype scheme of a linear power amplifier

The input signal is (in this case) fed into the transistor's base. This is not the only possible topology for power amplification. Some other topologies and their performance comparison will be shown in the next chapter. The DC bias voltage is also fed into the base. The collector inductor L_C allows the collector voltage to swing around the V_{DC} supply voltage. The maximum RF swing here available is then V_{DC} (since the collector voltage cannot exceed $2V_{DC}$). The aim of this “choke” inductor is also to allow only to the DC current to flow into it. The RF current will flow only into the load.

In the following paragraphs the collector current over base-emitter voltage characteristic is assumed to be as reported in Figure 2.2. This linear piecewise characteristic shows null current when the V_{BE} is lower than the threshold, while showing hard clipping when the V_{BE} reaches its maximum value $V_{BE,MAX}$.

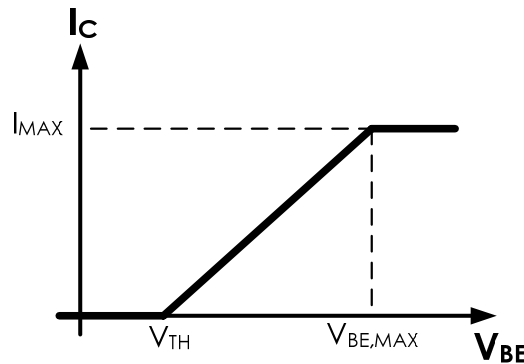


Figure 2.2 Ideal linear piecewise characteristic of a transistor

The maximum linear output power is reached at the 1-dB compression point (Figure 2.3). This is the value of the input power where the output power is reduced by one dB compared to the case of perfectly linear system. This behavior is due to the nonlinearities in the collector current characteristic which generate harmonics that reduce the linear power gain of the power amplifier. Referring to Figure 2.2, the 1dB compression point is reached when the collector current reaches the maximum value I_{MAX} where it is subject to hard clipping.

Since the collector voltage is limited by the supply voltage, the maximum linear output power will be determined by the R_{OUT} value, since it is linked to the collector voltage by the following equation:

2.1

$$P_{OUT} = \frac{V_{OUT}^2}{2R_{OUT}}$$

Thus, since the maximum swing is limited by the supply voltage, a low output impedance must be provided at the transistor's collector. For example, if a supply voltage of 3V is available, if one wants to deliver 1W to the load, a resistance of 4.5Ω has to be synthesized. This needs the use of an impedance transformation network, which properly reduces the 50Ω antenna impedance into the desired value. This matching network must also provide the reactive impedance which resonates out the transistor's output parasitic capacitance.

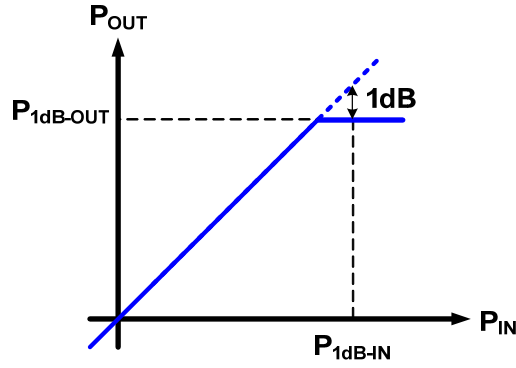


Figure 2.3 Compression point of an ideal amplifier

As already mentioned a power amplification transforms the DC supply power into RF power. The common way to measure how efficiently this conversion takes place is obviously named drain efficiency:

$$\eta = \frac{P_{RF}}{P_{DC}} \quad 2.2$$

which is the ratio between the RF power supplied to the load and the DC power supplied by the batteries. The amount of power supplied by the batteries which becomes heat is given by the difference between the DC power and the RF power:

$$P_{HEAT} = P_{DC} - P_{RF} \quad 2.3$$

Since a linear power amplifier shows the maximum efficiency at the maximum output power, the maximum heat dissipation may take place at reduced power levels where the efficiency is low (as will be shown in Appendix).

The power gain is defined by:

$$G_P = \frac{P_{RF}}{P_{IN}} \quad 2.4$$

This power gain should also be maximized, in order to relax the need of supplied RF power that the device which drives the amplifier (generally an up-conversion mixer) has to provide. Thus an efficiency which takes into account the power gain can be defined:

$$PAE = \frac{P_{RF} - P_{IN}}{P_{DC}} = \eta \left(1 - \frac{1}{G_P} \right) \quad 2.5$$

Named Power Added Efficiency, this figure of merit measures how much RF power is “added” to the input power. If the power gain is higher than 10dB the PAE approaches the drain efficiency.

As already mentioned, it is desirable to maximize the efficiency in a power amplifier. This can be obviously achieved by reducing the DC power supplied for a given RF output power. Since the supply voltage is fixed, the only way to increase the maximum efficiency is by reducing the DC current dissipated. This issue is addressed in the next paragraph.

2.2 Working Classes

Given a collector current I_C , the quiescent current I_{DC} dissipated by the transistor is given by:

$$I_{DC} = \frac{1}{T} \int_0^T I_C dt$$

This average current can be reduced by using a proper current shape with a low average value. For example a rectified sinusoid has a lower quiescent value compared to a full wave sinusoid.

It is possible to generate a collector current with a reduced DC current by properly biasing the transistor. The bias voltage determines the working class of the power amplifier, setting the conduction angle of the collector current. The concept of conduction angle is shown in Figure 2.4. It is defined as the portion of the RF signal where the current conduction in the transistor takes place. In the figure the input voltage and the collector current behavior are reported. It is possible to see that the conduction angle is reduced as the quiescent bias voltage (V_q) approaches the threshold of the transistor (V_t).

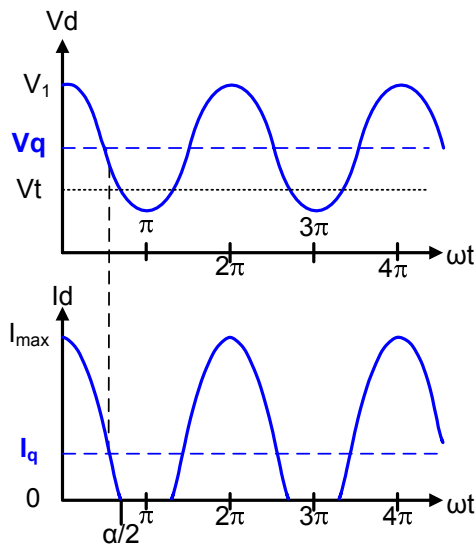


Figure 2.4 Conduction angle

Depending on the conduction angle it is possible to define several working classes, as reported in Table 2.1 [6].

| CLASS | Quiescent Voltage (V_q) | Quiescent Current (I_q) | Conduction angle (α) | Maximum Efficiency |
|-------|-----------------------------|-----------------------------|-------------------------------|--------------------|
| A | $\frac{1}{2}$ | $\frac{1}{2}$ | 2π | 50% |
| AB | 0 - $\frac{1}{2}$ | 0 - $\frac{1}{2}$ | $\pi - 2\pi$ | 50%-78.5% |
| B | 0 | 0 | π | 78.5% |
| C | <0 | 0 | 0 - π | >78.5% |

Table 2.1 Efficiency of the ideal working classes

A reduced conduction angle allows to reduce the quiescent current in order to decrease the DC power dissipated. Table 2.1 shows the well known results in terms of efficiency achieved by those power classes. It must be note that these efficiencies are the maximum achievable ones, which are obtained at the 1-dB compression point where the maximum current and voltage swing are reached. At lower power levels the efficiency decreases with a slope which depends on the amplification class of operation.

Before looking in detail how the conduction angle influences the shape of the efficiency, it is necessary to introduce the concepts of voltage and current efficiency. These concepts are not so widely used in the world of power amplifiers even if they are useful in order to understand the limitations of the various amplification classes. We should start with considering again the drain efficiency expression by replacing the RF and DC power with their explicit expressions as a function of voltage and current:

$$\eta = \frac{P_{RF}}{P_{DC}} = \frac{1}{2} \frac{V_{RF} I_{RF}}{V_{DC} I_{DC}} \quad 2.7$$

The coefficient $\frac{1}{2}$ arises from the sinusoidal nature of the signal considered. It is then possible to see that the drain efficiency can be split in two different efficiencies named voltage and current efficiency:

$$\eta_V = \frac{V_{RF}}{V_{DC}} \quad 2.8$$

$$\eta_I = \frac{1}{2} \frac{I_{RF}}{I_{DC}} \quad 2.9$$

It is now possible to analyze the efficiency behavior of the class A and class B power amplifier and then to make a comparison between them.

2.2.1 Class A

A well known result about linear power amplifiers states that the efficiency of a Class A power amplifier is 50%. This result just refers to the maximum achievable efficiency in ideal conditions. But when an amplitude modulated signal is considered it is evident that this condition represents a particular case during transmission. As already explained in Chapter 1, a modulated signal for the UMTS standards transmits with an average output power which is well below compared to the maximum value. Thus the efficiency must be maximized at the lower power levels.

Let's now consider the voltage and drain efficiency of a class A power amplifier as a function of the input signal. First, it must be note that the amplifier is biased with a fixed DC current I_{DC} which is not dependent on the input signal. Assuming that the transistor operates as a linear transconductor, the output instantaneous current is a linear function of the base voltage.

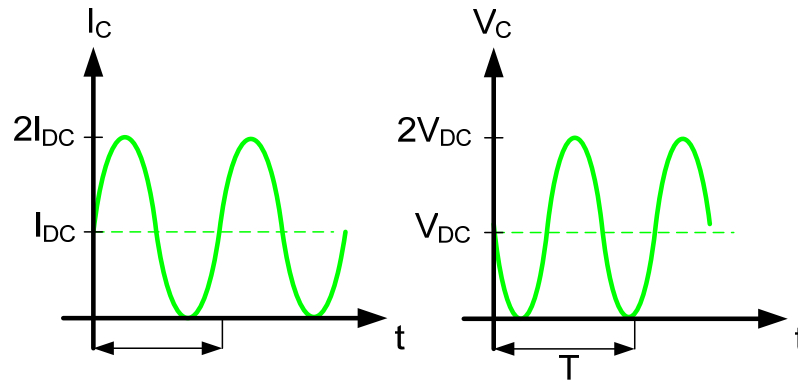


Figure 2.5 Class A PA waveforms

The current and voltage at the transistor's collector swing around the DC bias (Figure 2.5). The RF current is a function of the input signal, while the RF voltage is a function of the output current and the output resistance seen by the transistor. Since the collector RF current will have a maximum value of I_{DC} while the voltage has a maximum value of V_{DC} , at this point the voltage and current efficiency will be:

$$2.10 \quad \eta_{V,MAX} = \frac{V_{RF,MAX}}{V_{DC}} = \frac{V_{DC}}{V_{DC}} = 1$$

$$2.11 \quad \eta_{I,MAX} = \frac{1}{2} \frac{I_{RF,MAX}}{I_{DC}} = \frac{1}{2} \frac{I_{DC}}{I_{DC}} = \frac{1}{2}$$

thus giving the well known 50% maximum drain efficiency. But if we consider to have the amplifier operating at half of the maximum power, both the RF current and voltage will be reduced by $\sqrt{2}$, thus halving the drain efficiency to 25%. This is due to the fact that the quiescent current and voltage are fixed, while RF power changes dynamically. In Figure 2.6 the behavior of the voltage, current and drain efficiency are reported.

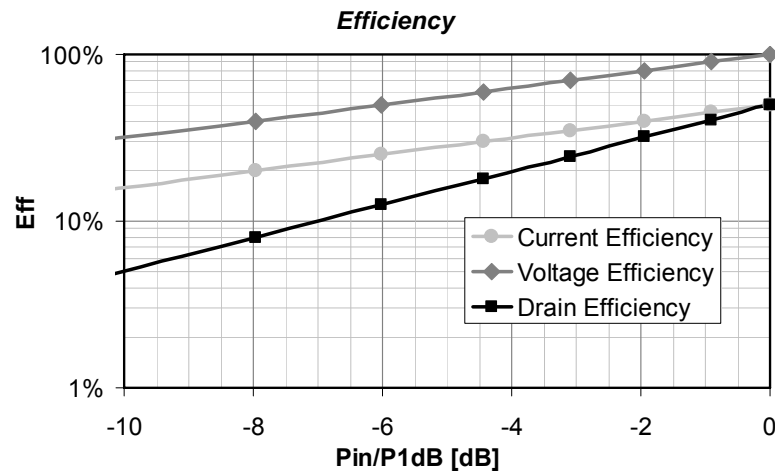


Figure 2.6 Class A PA Efficiencies

2.2.2 Class B

In the Class B power amplifier the transistor is biased at the threshold. This means that conduction takes place only for half of the input signal period. The DC current dissipated is then reduced compared to the Class A condition, where power is dissipated even if no signal is applied to the transistor. Thus the efficiency in a Class B is expected to be higher compared to the Class A, at least at the maximum RF power supplied. What happens when the output power is in backed-off conditions?

Let us first look at Figure 2.7, where the collector current and voltage are reported. From the voltage efficiency point of view, a Class B amplifier works as the Class A. This behavior is the same for all the linear classes. For what concern the current efficiency, we should calculate the RF and the DC current of a rectified sinusoidal current. Supposing to have this type of signal with a maximum value I_{MAX} , the RF and DC current will be:

$$I_{RF} = \frac{I_{MAX}}{2} \quad 2.12$$

$$I_{DC} = \frac{I_{MAX}}{\pi} \quad 2.13$$

Thus the current efficiency will be:

$$\eta_I = \frac{1}{2} \frac{I_{RF}}{I_{DC}} = \frac{\pi}{4} \quad 2.14$$

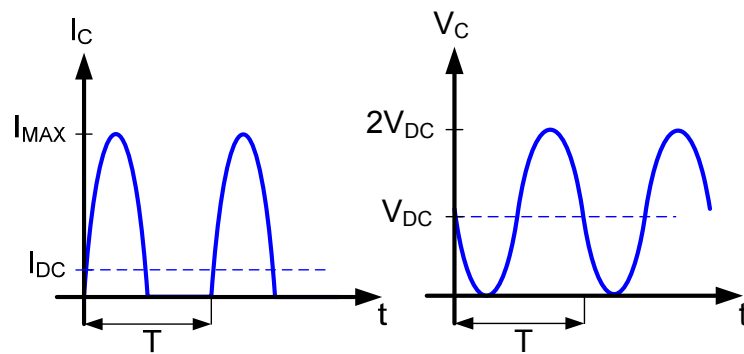


Figure 2.7 Class B PA waveforms

It is very important to note that this efficiency does not depend on the value of the signal applied: it is constant over the entire power range. The drain efficiency thus depends only on the voltage efficiency, which is maximum only at the 1dB compression point. The shapes of voltage, current and drain efficiency are reported in the next figure.

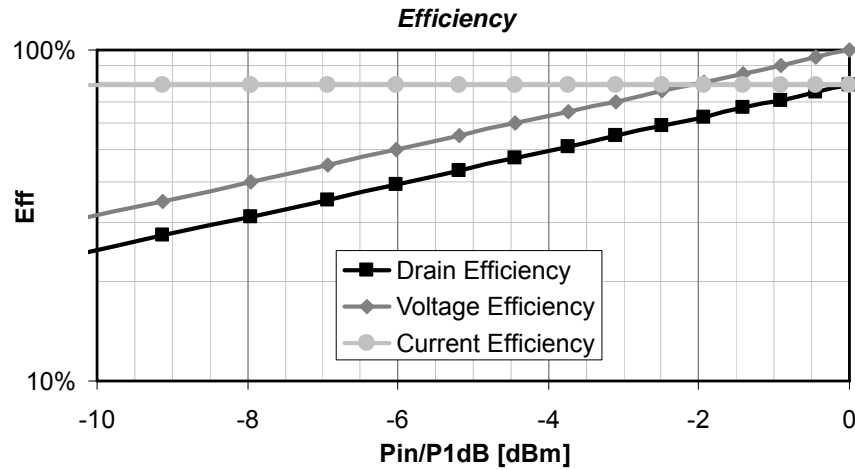


Figure 2.8 Class B PA Efficiencies

When the output power is halved compared to the maximum value, the efficiency drops from 78.5% to 55% which is a $\sqrt{2}$ factor. Thus, compared to the Class A condition, the efficiency reduction in the Class B is lower. This consideration is confirmed by the graph reported in Figure 2.9 which shows the ratio between the Class B and the Class A drain efficiency. It is then possible to see that the former has better performance compared to the latter not only in terms of maximum achievable efficiency, but also at the back-off.

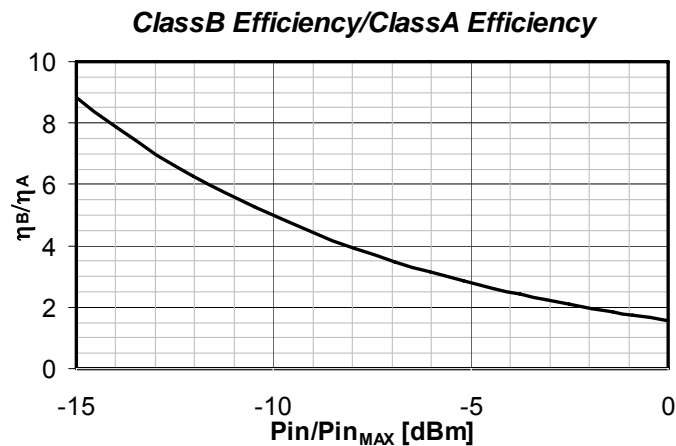


Figure 2.9 Ratio between Class A and Class B Efficiency

The better performance of a Class B Power Amplifier in terms of efficiency are not enough when employed with a UMTS modulated signal. In this case the average power is several dB below the maximum output power (typically from 6 to 11 dB under), where the efficiency is halved compared to the maximum value. It is then desirable to maximize the efficiency around this value. Since the current efficiency doesn't change in the overall power range, the only way to maximize the drain efficiency is related to the maximization of the voltage efficiency. This efficiency is maximum when the RF voltage swing equals the supply voltage V_{DC} . Since the voltage swing is related to the output resistance seen by the transistor, it is necessary to increase its value at the power level where the efficiency has to be maximized. At higher power levels it is possible to maintain the efficiency closer to the maximum value while dynamically

reducing it in order to keep the RF voltage swing closer to the maximum available.

This is the working principle of the Doherty technique, which tries to maximize the efficiency over a power range higher compared to a standard Class B amplifier. This technique will be shown later in this chapter, and compared to some other efficiency enhancement techniques.

2.2.3 Class AB

In the real world a “pure” class B cannot exist. This is because the actual transcharacteristic of a transistor is not a linear piecewise function. Especially around the threshold voltage, the current doesn’t show an abrupt change.

A more linear but still efficient solution is to bias the transistor in a Class AB condition. As the name suggests, this is a amplification class “in between” class A and Class B. It is possible to show that such a power amplifier is actually sufficiently linear while still showing good efficiency. In this class of amplification the transistor is biased slightly above the threshold voltage. With a small signal applied, the conduction angle is the same as in a Class A amplifier. When the input signal increases, the conduction angle starts to be reduced, until the 1dB compression point is reached and the collector current shape is similar to that of a Class B.

It is intuitive that the voltage efficiency has the same behavior as in the other amplification classes, while the current efficiency is different than both. In fact the current efficiency will be similar to that of a Class A amplifier when the signal applied is small, becoming almost constant at the higher power level. Thus, the advantage of a Class B amplifier is maintained at the higher power level, while at the lower power level the poor efficiency performance does not impact the overall efficiency in a dramatic way.

2.2.4 Harmonics generation

The reduction of the conduction angle allows to reduce the DC component of the drain/collector current. This is obtained avoiding conduction in the transistor and then modifying the sinusoidal current shape. This “truncation” of the current introduces some high frequency components which amplitude increases with the conduction angle reduction. Figure 2.10 [6] shows the various harmonics generated with different conduction angle. It is possible to note that (ideally) throughout the class AB range and up to the midway class B condition the only significant harmonic, other than the fundamental, is the second. Thus, in order to have a correct current shape which implements the class AB or B it is necessary to provide a low impedance path to the second harmonic current, otherwise the current will have the form of a sinusoid whatever the bias condition is. This wouldn’t allow to have a reduced conduction angle current, thus limiting the efficiency.

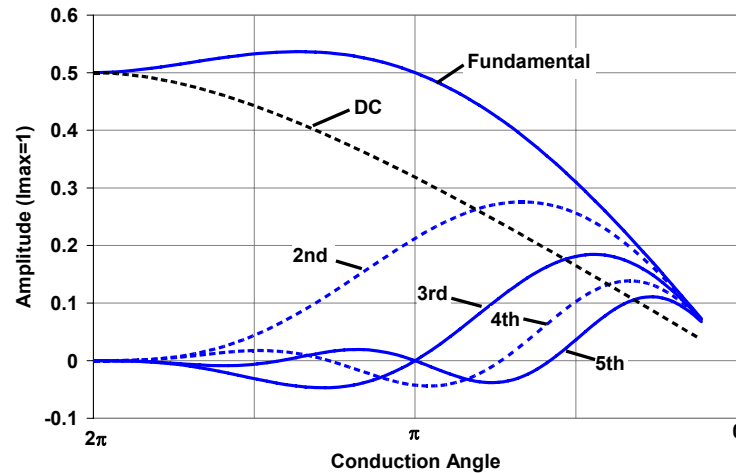


Figure 2.10 Harmonics of the drain current generated by a reduced conduction angle

2.3 Efficiency Enhancement Techniques

As already mentioned, when an amplitude modulated signal is considered, the efficiency must be maximized in the backed-off conditions. This task can be achieved in different ways, each of them has the relative advantages and disadvantages. In the next paragraphs several efficiency enhancement techniques will be introduced. The design of a Doherty linear amplifier will be shown in detail in Chapter 4.

Several efficiency enhancement techniques will be shown in the next paragraphs. These will be the Doherty architecture, Envelope Tracking, Chireix outphasing and Envelope Elimination and Restoration (EER). These techniques are based on different way to maximize the efficiency while trying to maintain the linearity in the power amplification. The theoretical performance of these techniques are reported in Figure 2.11[7]. It has to be note that in the real world these performance suffer of reduction due to the drawback that each technique carries with itself.

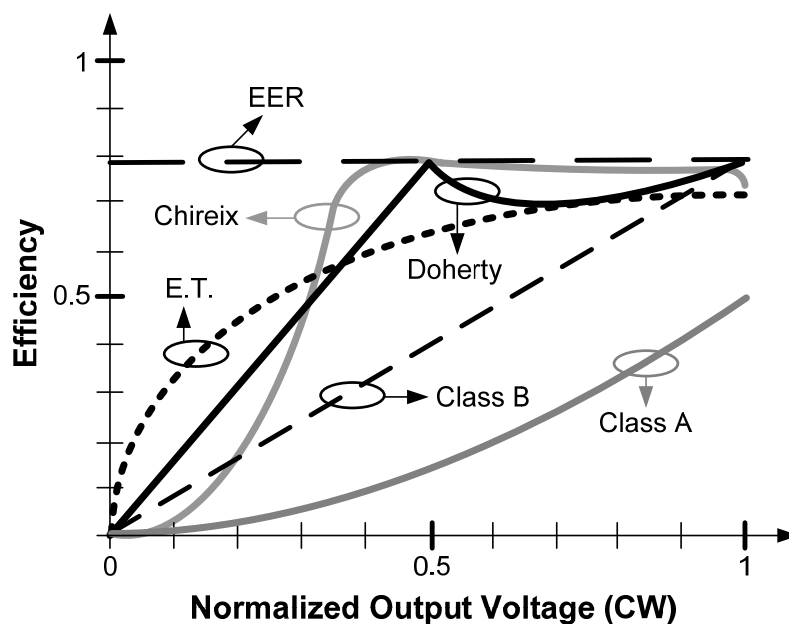


Figure 2.11 Performance of several efficiency enhancement techniques

2.3.1 Doherty Power Amplifier

This efficiency enhancement technique has been introduced by W.H. Doherty in 1936 [9]. At that time the broadcast radio communications employed AM modulated signal. The continuous enhancement of the power level used by the broadcasting devices created several thermal issues due to dissipated power. Since Class A power amplifier were used (in order to have a good linearity) the limited efficiency at the reduced power level made difficult to cool down the vacuum tube used.

The basic idea of the Doherty power amplifier is to maximize the voltage efficiency by a convenient variation of the output resistance. This makes the overall efficiency to be close to its maximum value in the range where the resistance is dynamically varied at the lower power levels. Supposing to have a Class B power amplifier, if the output resistance which maximizes the efficiency (maximizing η_v) at the 1dB compression point (P_{1dB}) is R , if one wants to maximize the efficiency at, that say, a quarter of the P_{1dB} , the resistance that must be seen at the collector should be $4R$ (because the RF voltage is reduced by a factor of 4). In order to keep the efficiency close to the maximum value in the range between $0.25P_{1dB}$ and P_{1dB} it is necessary to dynamically reduce the resistance seen at the collector from $4R$ to R . A possible way to reduce dynamically the impedance seen at the output of an amplifier is shown in Figure 2.12.

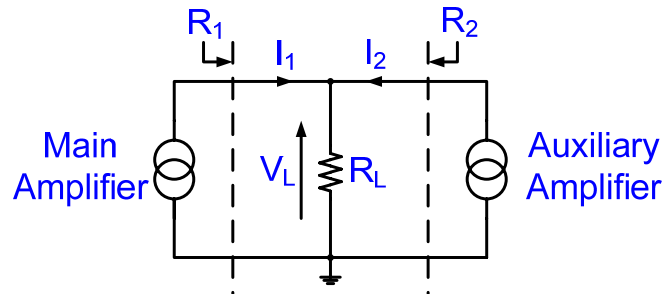


Figure 2.12 Dynamic resistance variation principle

Here the power amplifier is modeled by an ideal current source (named Main Amplifier). When the Auxiliary is off, the impedance seen at the Main Amplifier Output (R_1) equals R_L . When the two amplifiers are both on, the voltage across R_L is:

$$V_L = R_L (I_1 + I_2) \quad 2.15$$

While R_1 is:

$$R_1 = \frac{V_L}{I_1} = R_L \left(\frac{I_1 + I_2}{I_1} \right) \quad 2.16$$

At the same time the Auxiliary amplifier “sees” an output resistance:

$$2.17 \quad R_2 = \frac{V_L}{I_2} = R_L \left(\frac{I_1 + I_2}{I_2} \right)$$

It is then possible to vary the resistance seen at one amplifier due to the effect of the current generated by the other one. Generally, if the two current supplied by the amplifier have different phases, the impedance seen at the main amplifier will be:

$$2.18 \quad \bar{Z}_1 = R_L \left(1 + \frac{\bar{I}_2}{I_1} \right)$$

If I_1 and I_2 are in phase, Z_1 can be transformed to higher resistive values, while if the two currents have opposite phases, the output impedance seen by the main amplifier can be reduced. In a real case this scheme has to be changed in order to generate the effective efficiency enhancement. The design of a real Doherty Power Amplifier will be discussed later in Chapter 4.

The drain efficiency of a Doherty power amplifier compared to a single stage class B PA is shown in Figure 2.11. The efficiency is maximized at a quarter of the P_{1dB} and allows to gain a 50% of the efficiency compared to the Class B PA. This gain in efficiency takes place also at the lower power levels, while in the upper range the average efficiency is still higher. This behavior makes the Doherty solution very attractive, showing a great benefit in the overall efficiency. However this is the ideal behavior. In the real world some effect will reduce the efficiency of a Doherty PA, since a tradeoff between linearity and efficiency has to be taken into account. This effects will be studied in Chapter 4.

2.3.2 Envelope Tracking (Bias Adaptation)

As already discussed, the aim of a Doherty PA is to maximize the efficiency of a Class B PA while maximizing its voltage efficiency. This result is achieved by a convenient variation of the output impedance, which keeps the amplifier to operate with a full-swing output voltage.

This is not the only way to maximize the voltage efficiency. In fact, since it is given by the ratio between the RF voltage and the DC supply voltage at the PA's output, it is possible to maximize it with a variation of the DC supply voltage when the amplifier works at lower power levels. The principle scheme which implements this operation is shown in Figure 2.13. Here an envelope detector reveals the amplitude variation in the input signal, giving a DC voltage proportional to the envelope variations. This variable “small signal” DC voltage is boosted by a DC-DC converter to a variable supply voltage at the power amplifier. It must be note that the power amplifier used is still a linear power amplifier.

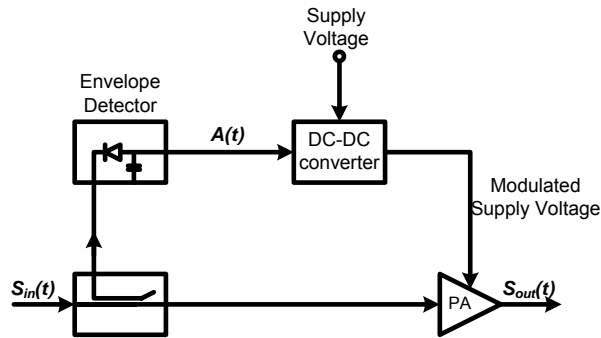


Figure 2.13 Envelope tracking principle

In this way the voltage efficiency is kept close to the maximum value, thus increasing the drain efficiency. A real example [8] of dynamic V_{DD} modulation is reported in Figure 2.14 a), while the efficiency gain achieved is shown in Figure 2.14 b). The supply voltage is kept constant when the RF voltage (and then the RF power) is below a certain breakpoint, where the efficiency is wanted to be maximized. When the RF power higher than at the breakpoint, the supply voltage is dynamically changed, increasing the efficiency compared to the fixed-voltage solution.

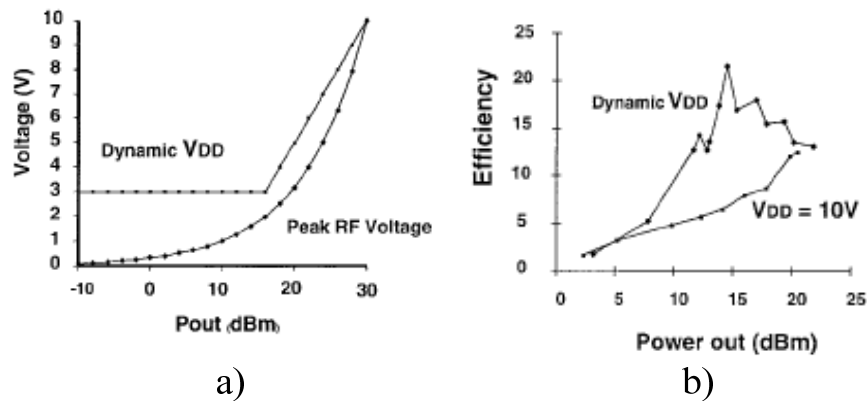


Figure 2.14 Example of bias adaptation performance

This efficiency enhancement technique employs the voltage efficiency maximization idea in an alternative way compared to the Doherty solution. Since the amplitude modulation of the input signal it is not affected by the supply boosting system (since a linear PA is used), there aren't particular constraint in the envelope detector, since it doesn't need to generate a precise voltage variation. The drawback of this technique is related to the efficiency of the DC-DC conversion, which affects the overall efficiency of the power amplifier.

2.3.3 Chireix Amplifier (LINC)

In the solutions shown above, the efficiency of a linear PA is increased by maximizing its voltage efficiency. The technique here discussed addresses the problem of efficiency enhancement from a different perspective. A non linear and efficient power amplifier is used but the output AM modulated signal doesn't (ideally) show distortion. This technique was born in the same period of

the Doherty technique. Like the Doherty two amplifiers are employed, but in this case they are both not linear.

The main difference between a non linear power amplifier and a linear one arises from the way the transistors are used. In a linear PA the transistor acts as a transconductor, while in a non linear amplifier the transistor is used as a switch. In this way the amplifier is inherently more efficient, but its natural application is related to the amplification of the constant envelope signal. How it is possible to linearly amplify a modulated signal by using non linear switching amplifiers [9] ?

Let's first consider this trigonometric relation:

$$2.19 \quad \cos(A) + \cos(B) = 2 \cos\left(\frac{A+B}{2}\right) \cos\left(\frac{A-B}{2}\right)$$

If the angles A and B are variable in the time domain ($A = \omega t + \phi$ and $B = \omega t - \phi$), the previous equation can be rewritten as:

$$2.20 \quad \cos(\omega t + \phi) + \cos(\omega t - \phi) = 2 \cos(\phi) \cos(\omega t)$$

Let's now consider Figure 2. 15. If an amplitude modulated signal $S_{IN}(t) = A(t)\cos(\omega t)$ is applied to a phase modulator it is ideally possible to generate two signals $S_1(t)$ and $S_2(t)$ with a fixed amplitude and a phase depending by the amplitude of the input signal:

$$2.21 \quad S_1(t) = \cos\{\omega t + \cos^{-1}[A(t)]\}$$

$$2.22 \quad S_2(t) = \cos\{\omega t - \cos^{-1}[A(t)]\}$$

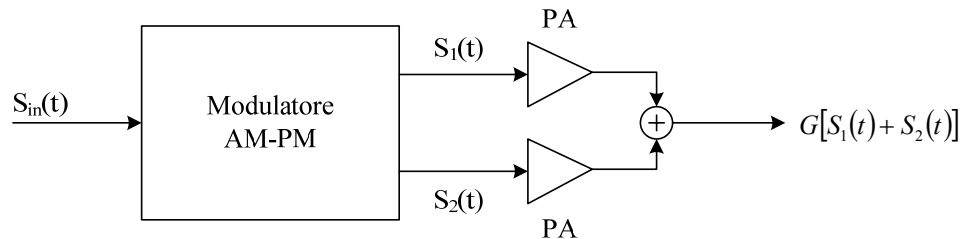


Figure 2. 15

If the PAs used have a voltage gain G and $\phi = \cos^{-1}[A(t)]$ the sum of the two signals gives (applying the 2.20):

$$2.23 \quad S_{out}(t) = G[S_1(t) + S_2(t)] = 2GA(t)\cos(\omega t)$$

Which is an amplified replica of the input signal. The key element in a LINC structure is the phase modulator, which converts the amplitude modulation in two phase modulated signals with a phase shift of 180° . It has to be note that the amplifier used can be nonlinear, since the two signal $S_1(t)$ and $S_2(t)$ doesn't carry information in their amplitude. The amplitude modulation is restored at the

output by the sum of the two signals, which doesn't affect the phase variation in the signals.

Summation of the out-of-phase signal in a nonhybrid linear combiner inherently results in a variable reactive PA-load impedances. If the combiner is untuned, the current drawn from the PAs is proportional to the transmitter-output voltage, resulting in efficiency characteristic that varies with signal amplitude, as in a similar Class B PA. The Chireix technique uses shunt reactances on the inputs of the combiner (Figure 2.16) to tune out the drain reactances at a particular amplitude, which, in turn, maximizes the efficiency in the vicinity of that amplitude. A proper choice of the shunt susceptances, the average efficiency can be maximized for any given signal.

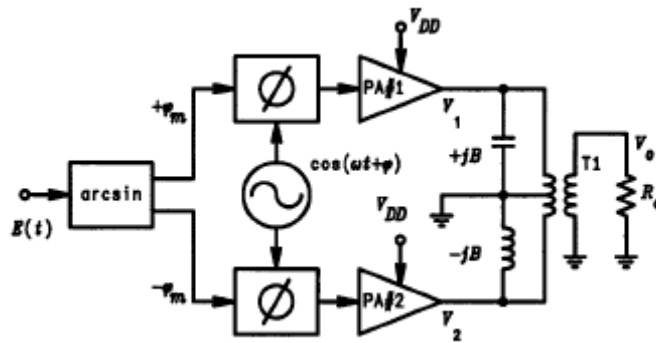


Figure 2.16 Chireix amplifier

2.3.4 Envelope Elimination and Restoration

This technique, originally proposed by Kahn [11], combines the use of a nonlinear but efficient PA with an envelope amplifier, in order to obtain a linear and efficient amplifier. The classic implementation of this technique a voltage limiter cancels out the amplitude variations at the signal applied to the power amplifier, maintaining only the phase variations of the input signal. The signal envelope can be restored at the output via a modulation of the supply voltage.

The high-level AM modulation process is in principle more efficient compared to a traditional linear amplification approach. Since the amplifier used is not linear, its efficiency is not affected by variations in the supply voltage. Then the amplitude of the envelope at the RF out will be proportional only to the supply modulation.

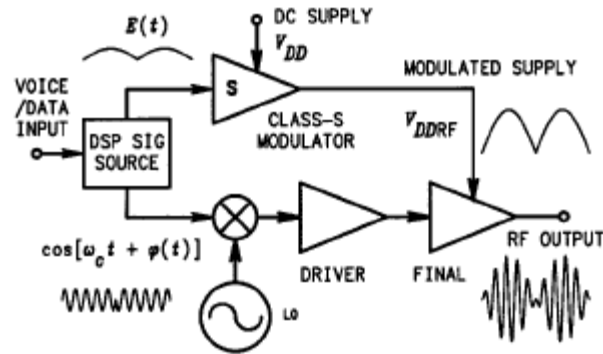


Figure 2.17 EER amplifier

The transmitters based on this technique have theoretically an excellent linearity since it depends only by the amplitude modulator (which works at a lower frequency) and not by the power amplifier. The major effects which limit the linearity are related to the envelope bandwidth and the alignment between amplitude and phase modulation. The envelope bandwidth has to be at list the double of the RF signal bandwidth. Moreover, the misalignment between the amplitude and phase modulation must not exceed the 10% of the inverse in the RF bandwidth. These constraint has limited the wide use of this technique for limited bandwidth signals (like GPRS and EDGE).

2.4 Conclusion

In this chapter the main characteristics of a linear power amplifier has been presented. The important concept of voltage and current efficiency allows to better understand the ways to increase the efficiency of a power amplifier. Several efficiency enhancement techniques have been presented. The Doherty and Bias adaptation goal is to maximize the voltage efficiency of a linear power amplifier in order to maximize the overall efficiency on a wider power range. The Chireix and EER techniques employ non linear power amplifiers and their goal is to efficiently amplify a phase modulated signal an restore the amplitude variations at the output. The Doherty technique seems to be the most straightforward way to enhance the efficiency of a linear power amplifier, since it is not band limited and the linearity is determined only by the main amplifier.

References

- [6] Steve C. Cripps, “RF Power Amplifiers for Wireless Communications”, *Artech, 1999*
- [7] F.H. Raab et al. “*Power Amplifiers and Transmitters for RF and Microwave*”, IEEE Transactions on Microwave Theory and Techniques, Vol. 50, No. 3, March 2002
- [8] Hanington, G.; Pin-Fan Chen; Asbeck, P.M.; Larson, L.E.; “*High-efficiency power amplifier using dynamic power-supply voltage for CDMA applications*” Microwave Theory and Techniques, IEEE Transactions on Volume 47, Issue 8, Aug. 1999 Page(s):1471 - 1476
- [9] W. H. Doherty, “*A new high efficiency power amplifier for modulated waves,*” Proc. IRE, vol. 24, pp. 1163–1182, Sept. 1936
- [10] R. Cannizzaro, “Trasmettitore multistandard LINC integrato in tecnologia CMOS: studio architetturale e progetto dell'amplificatore e del ricombinatore”, Pavia, A.A. 2003-2004
- [11] Kahn, L.R.; “*Single-Sideband Transmission by Envelope Elimination and Restoration*” Proceedings of the IRE Volume 40, Issue 7, July 1952 Page(s):803 - 806

Chapter 3

A Common Base Class AB PA Design and Testing

This Chapter will deal with the design of a class AB linear power amplifier in a Si:Ge 0.25 μ m technology. The amplifier is based on a common base topology and employs a resonant network for passive current amplification. This allow to theoretically increase the efficiency of a standard cascode amplifier. Moreover this topology allows to bias the output stage with a relatively high voltage (beyond BVCEO) since it is insensitive by the avalanche breakdown current generated by the bipolar transistor (this issue will be introduced in the first paragraph). A performance comparison among different amplification topology will be discussed, posing the attention to the benefits introduced by a passive current amplification. The complete design flow will be shown, and the measurement on its realization will be discussed. In the last part of the chapter thermal issues regarding the test board and their effect on the PA performance will be addressed.

3.1 Breakdown Mechanism in Bipolar Transistors

A bipolar transistor is made by two p-n junctions, so it suffers from avalanche breakdown current generation.

When an electric field is applied to a reverse biased p-n junction a small reverse current of minority carriers flows from the n to the p region through the depletion region. When the reverse bias voltage is increased, this current also increases: if a critical electric field is applied, the carriers traversing the depletion region acquire sufficient energy to create a new hole-electron pair in collision with silicon atoms. This is called the avalanche process and leads to a sudden increase in the reverse-bias leakage current since the newly created carriers are also capable of producing avalanche [13].

The avalanche current near the breakdown voltage in a p-n junction can be written as:

$$I_{RA} = I_R M \quad 3.1$$

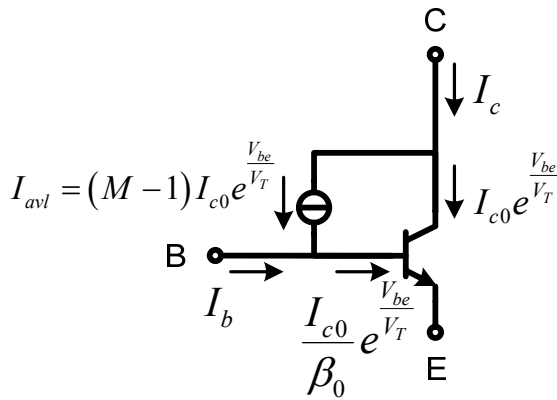
where I_R is the normal reverse bias current with no avalanche effect and M is the multiplication factor defined by

$$M = \frac{1}{1 - \left(\frac{V_R}{BV}\right)^n} \quad 3.2$$

In this equation, V_R is the reverse bias applied to the junctions, BV is the breakdown voltage and n has a value between 3 and 6.

For what concern a bipolar transistor, two BV can be defined: BV_{CBO} (collector-base breakdown with the emitter open) and BV_{CEO} (collector-emitter breakdown with base open).

The first breakdown effect is the same as the p-n junction breakdown, since the CB junction is reverse biased in a bipolar transistor in the active region. The model that includes the avalanche current effect (valid in the forward active region) is shown in Figure 3.1.



The collector current expression follows the 3.1:

$$I_c = M \cdot I_{c0} \cdot e^{\frac{V_{be}}{V_T}} \quad 3.3$$

where V_{be} is the voltage across the base-emitter junction and $V_T = kT/q$ is the thermal voltage.

Figure 3.1 *Avalanche current model*

The base current is the difference between the collector current (scaled by the transistor's DC current gain, β_0) and the recombination current:

$$3.4 \quad I_b = \frac{I_{c0}}{\beta_0} e^{\frac{V_{be}}{V_T}} - (M-1) I_{c0} e^{\frac{V_{be}}{V_T}}$$

In this case the avalanche current multiplication factor is defined as:

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

As V_{cb} approaches BV_{CBO} , the avalanche multiplication factor $M \rightarrow \infty$ and the collector-base junction breaks down. This breakdown phenomenon is independent by the impedance connected between the base and emitter terminals. Therefore, BV_{CBO} is an absolute maximum for V_{cb} , and circuits should be designed to operate at $V_{cb} < BV_{CBO}$ under all operating conditions. Typical values of BV_{CBO} range from around 10 to 16 V, which are sufficiently high but that still can be instantaneously reached in handsets applications (i.e. in case of antenna disconnection during transmission).

The collector-emitter breakdown effect takes place when the recombination current (last term in 3.4) nulls the base current. In this case, after handling the 3.4 it is possible to find the BV_{CEO} expression:

$$BV_{CEO} = V_{be} + \frac{BV_{CBO}}{\sqrt[n]{\beta_0 + 1}}$$

This breakdown mechanism is quite different from the previous one, since it doesn't work just like a p-n junction breakdown. This is because hole-electron pairs are produced by the avalanche process and the holes are swept into the base, where they effectively contribute to the base current. In a sense the avalanche current is then amplified by the transistors since the recombination current hasn't a low impedance path to flow in (the base is open). In the other breakdown case the base was considered shorted, and the effect of the base current amplification was not an issue.

The 3.6 shows that BV_{CEO} is a few times lower than BV_{CBO} (around 3 times) and it can be considered as an inferior limit to the breakdown voltage. Anyway this limit can be overcome by supplying a low base impedance to the transistor on order to move the breakdown limit to the BV_{CBO} by using a proper bias network. This makes some bias configuration able to let the transistor's collector to be biased around the BV_{CEO} . This has the advantage to increase the PA's voltage efficiency since the output swing can be significantly high compared to the saturation voltage. Moreover, given a maximum deliverable output power, the increase of the collector supply voltage allows to relax the specification about the impedance transformation network which can be more efficient.

This strength to voltages higher than $BC_{V_{EO}}$ is also required due to reliability issues. In fact, load mismatch conditions, e.g. due to antenna impedance variations, can lead to instantaneous over-voltages that, for a voltage standing-wave ratio (VSWR) of 10:1, can go up to four times the supply voltage [14]. Under worst-case conditions, such as during battery recharge, an output voltage as high as 20 V may be reached [14]. This means that, even using high-voltage devices, special protection countermeasures need to be taken to avoid device failure.

An example of the base impedance effect is shown in figure 2, where the collector current to the V_{CE} of an ST Si:Ge 0.25 μ m transistor is shown under different base impedance conditions (respectively base open and shorted).

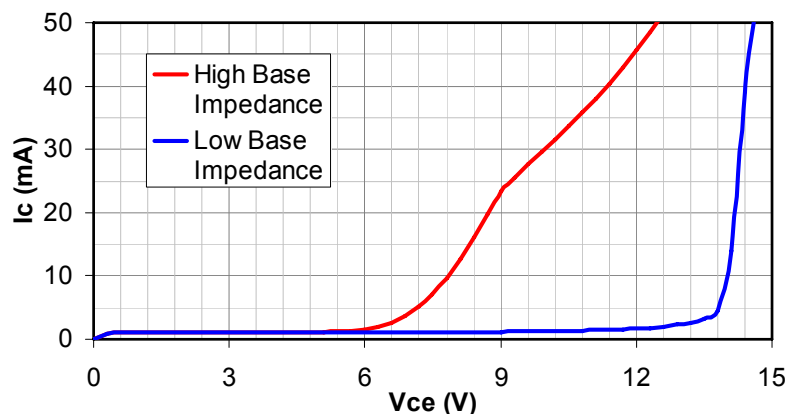


Figure 3.2 Effect of base impedance to the collector current

Looking at the graph it is possible to find the BV_{CEO} (6 V in the red curve) and the BV_{CBO} (13V in the blue curve). This result is partially in contrast with the previous statement of BV_{CEO} three times smaller than BV_{CBO} . This discrepancy is due to the fact that the 3.6 considers the plane junction breakdown, neglecting edge effects. This is

because it is only collector-base avalanche current actually under the emitter that is amplified, but the measured value of BV_{CBO} is usually determined by avalanche in the curved region of the collector, which is remote from the active base.

3.2 Basic stages comparison

The paragraph 3.1 has shown that a good strength against over-voltages is desirable, not just for reliability purposes, but also because it is possible to maximize the amplifier drain efficiency. Some basic amplification stages will now be discussed and their performance compared. The technology considered is the ST Si:Ge 0.25 μm which breakdown characteristics were previously shown in Figure 3.2.

3.2.1 Common Emitter

The common emitter PA is the classical solution adopted in the power stage of linear power amplifiers such as the one reported in Figure 3.3. The input signal is fed into the base while the output is taken at the collector. The L_C inductor allows the collector voltage to swing over V_{CC} (which is the collector supply voltage). Ideally, the maximum linear RF output voltage has an amplitude of $V_{CC} - V_{SAT}$ (where V_{SAT} is the minimum collector-emitter voltage that keeps the device in the linear operating region).

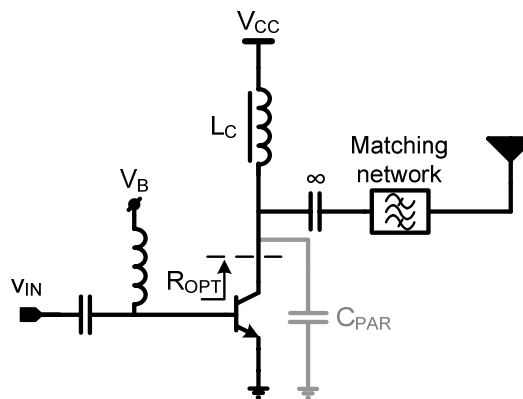


Figure 3.3 Common Emitter PA Stage

A proper impedance transformation network gives the optimum output load in order to have power amplification. The optimum load is given by a parallel inductance that cancels out the total output capacitance seen at the transistor collector and a parallel resistance that is set essentially by the supply voltage and the desired power level. Given a maximum power level of P_{SAT} , the optimum output resistance is:

3.7

$$R_{OPT} \cong \frac{(V_{CC} - V_{SAT})^2}{2P_{SAT}}$$

An high supply voltage is desirable since it allows to deliver a large output power with a relatively high optimum load impedance, simplifying the output impedance transformation network. However the choice of V_{CC} must be take into account the breakdown voltage characteristics of the technology used, for the motivations stated in paragraph 3.1. In this amplifier topology the DC bias network and the input signal are connected through the same terminal (i.e. the base), and it is not possible to supply a low impedance at DC and all other harmonics of the signal. Thus the maximum reachable collector voltage is the BV_{CEO} , since it is not possible to give a low impedance path to the recombination base current. For the considered technology, referring to Figure 3.2, this voltage is 6 V. This means that V_{CC} should not be chosen higher than 2.5V (in order to have some margin). Moreover, since the saturation voltage is around 1V, if a $P_{SAT} = 0.25W$ is chosen, the optimum output resistance will be:

$$R_{OPT} = \frac{(2.5-1)^2}{2 \cdot 0.25} \cong 4.5\Omega$$

which is a quite small value and makes the transformation network difficult to realize.

The transistor's size must be chosen in order to give the convenient output current I_{OUT} . At the maximum power the output voltage is $V_{OUT} = V_{CC} - V_{SAT} = 1.5V$ and the maximum output current will be:

$$I_{OUT} = \frac{2P_{SAT}}{V_{OUT}} = \frac{0.5}{1.5} \cong 330mA \quad 3.8$$

If the transistor is supposed to be biased in a class B condition, the quiescent current I_{DC} will be $\pi/2$ smaller than I_{OUT} . It is then possible to handle the efficiency expression maxing it depending only by the voltage efficiency:

$$\eta = \frac{P_{OUT}}{P_{DC}} = \frac{1}{2} \frac{I_{OUT} V_{OUT}}{I_{DC} V_{CC}} = \frac{\pi}{4} \frac{V_{OUT}}{V_{CC}} = \frac{\pi}{4} \left(1 - \frac{V_{SAT}}{V_{CC}} \right) \cong 47\% \quad 3.9$$

The [3.9 shows that the reduction of V_{CC} compared to V_{SAT} lowers the maximum PA efficiency, since V_{SAT} is a technology parameter.

The supply voltage can be augmented in two ways:

- Increasing the BV_{CEO} via new technology solutions while biasing the transistor below this limit;
- Using proper topologies/bias networks which allow to bias the transistor above BV_{CEO} ;

The first solution is in contrast with the actual technology evolution which moves toward higher frequency capabilities and contextually decreases breakdown voltage. At the technological level this has been addressed by providing several transistor designs, optimized for high power or high frequency operation [13]. Hence, using high-voltage transistors biased well below BV_{CEO} is a viable, albeit inefficient, solution under optimum load conditions.

The second is a more efficient solution and its capabilities will be shown in the next paragraphs.

3.2.2 Cascode

With a cascode topology it is possible to overcome the BV_{CEO} limit. This happens because the cascode transistor can be biased with a low impedance on its base: this gives a path to the avalanche current, as previously discussed. The ideal cascode PA is shown in Figure 3.4. The big capacitance C_{fat} gives the low impedance at the harmonics of the signal, while the bias network (which is not shown for simplicity) gives a low DC impedance. The bias networks must be also compliant with the avalanche current generated from the transistor. This matter will be discussed later.

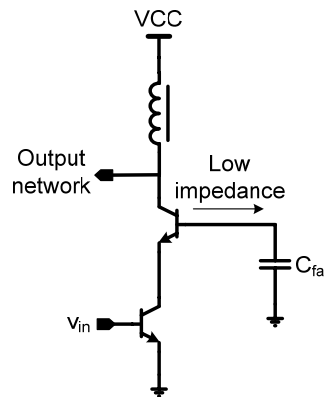


Figure 3.4 Ideal Cascode PA

As previously stated, this topology has the advantage to give a relatively high power supply, relaxing the output network design. Using the same technology of the example for the common emitter stage, it is now possible to bias the amplifier at $V_{CC} = 4.5\text{V}$ (which is less than half of BV_{CBO}). Then, for a $P_{SAT} = 0.25\text{W}$, the optimum output resistance will be:

$$3.10 \quad R_{OPT} = \frac{(V_{CC} - 2V_{SAT})^2}{2P_{SAT}} = \frac{(4.5 - 2)^2}{2 \cdot 0.25} = 12.5\Omega$$

which is relatively simple to achieve with a low-Q passive elements. Note that in 3.10 the output swing is now $V_{OUT} = V_{CC} - 2V_{SAT}$ because of the stacking of two transistor. This is a drawback for this topology since the additional power dissipated by Q_2 reduces the overall efficiency which has the form:

$$\eta = \frac{P_{OUT}}{P_{DC}} = \frac{\pi V_{OUT}}{4 V_{CC}} = \frac{\pi}{4} \left(1 - \frac{2V_{SAT}}{V_{CC}} \right) \cong 44\% \quad 3.11$$

The advantage of an higher supply voltage doesn't carry an improvement on the drain efficiency, because of the extra-power dissipated. However the reliability needs are satisfied, because this stage can sustain an higher output voltage compared to the common emitter stage.

The efficiency of this stage can be increased founding a way to decouple the two transistor, in order to loose just one V_{SAT} on the output voltage.

Some words should be spent about the output current; since the power supply (and so the output swing) has been increased and the target output power hasn't been changed, the maximum output current is reduced:

$$I_{OUT} = \frac{2P_{SAT}}{V_{OUT}} = \frac{0.5}{2.5} = 200mA \quad 3.12$$

This means that the transistor's size used in this solution is smaller than in the CE case, but the overall active area needed is anyway bigger.

3.2.3 Common Base

A possible way to exploit the advantage of an higher supply voltage without the stacking due to the cascode is to use a common base topology, shown in Figure 3.5.

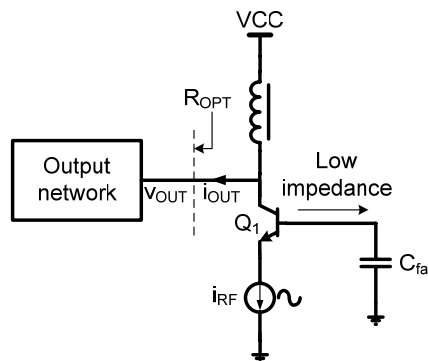


Figure 3.5 Ideal Common Base PA

In this case the input signal is assumed to be a current instead of a voltage. The utility of this choice will be more evident later, anyway an input emitter voltage can be also assumed. This stage is a current buffer so that the power amplification is due to the voltage amplification between the emitter and the collector. Also, a low impedance path can be furnished at the base, and this gives the possibility to increase the supply voltage V_{CC} . Moreover, the maximum output voltage swing is the same as in the common emitter topology: $V_{OUT} = V_{CC} - V_{SAT}$.

Thus, if the same assumption made for the cascode topology regarding V_{CC} , P_{SAT} and V_{SAT} are made, the optimum output resistance will be:

$$3.13 \quad R_{OPT} = \frac{(V_{CC} - V_{SAT})^2}{2P_{SAT}} = \frac{(4.5 - 1)^2}{2 \cdot 0.25} = 24.5\Omega$$

which is bigger than in the previous discussed case, because of the increased voltage swing. Also an higher efficiency is expected:

$$3.14 \quad \eta = \frac{P_{OUT}}{P_{DC}} = \frac{\pi V_{OUT}}{4 V_{CC}} = \frac{\pi}{4} \left(1 - \frac{V_{SAT}}{V_{CC}} \right) \cong 61\%$$

This result is more near to the 78.5% ideal maximum efficiency for a class B PA and gives a large improvement compared to the previous discussed solutions. Thus this solution seems to be more appealing since combines the advantages of both the common emitter and the cascode topologies. Moreover, a lower RF current is needed for this solution because of the higher voltage swing:

$$3.15 \quad I_{OUT} = I_{RF} = \frac{2P_{SAT}}{V_{OUT}} = \frac{0.5}{3.5} = 143mA$$

This results shows that a smaller transistor size can be used in this case giving an apparent saving of area. However, because of the current buffer characteristic of this topology, the i_{RF} must be supplied by the previous block in the transmitter chain, which is the upconversion mixer. Unfortunately (especially for high power application) this block is not able to provide an output current so high and this makes the use of a driver stage mandatory.

3.2.4 AC-Coupled Cascode

A Possible way to implement the driver stage in the common base topology is to AC-couple a common emitter (CE) driver stage and a common base (CB) output stage. This solution works like an AC-Coupled cascode, where the transistor Q_1 and Q_2 are connected with a big capacitance, as shown in Figure 3.6.

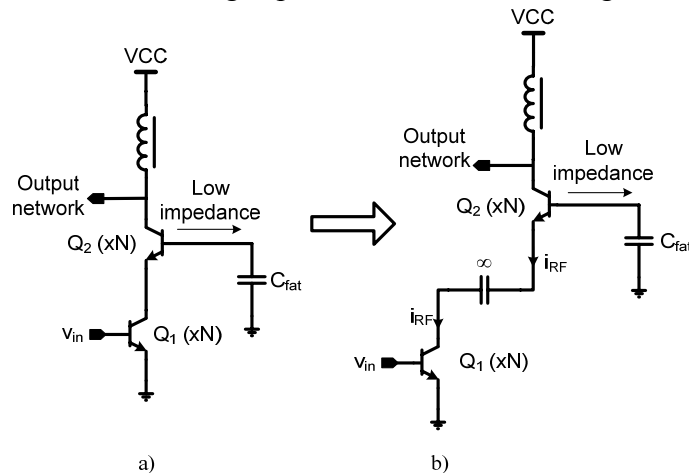


Figure 3.6 a) cascode amplifier; b) AC-coupled cascode.

However, even if the output stage has the advantages of a CB solution, the overall circuit has the same performance of a cascode topology, because of the additional power dissipated by Q_1 . A more efficient solution, conceptually similar to the one proposed in [12] is shown in Figure 3.7.

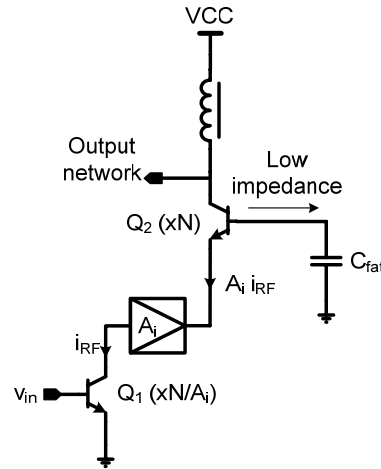


Figure 3.7 Cascode Amplifier with inter-stage impedance-transformation network.

The inter-stage passive impedance transformation network scales down the impedance level going from the common-emitter to the common-base device. This amounts to a passive current amplification and, as a result, the common-emitter device signal current level is significantly lower compared to the common-base device. Assuming that the inter-stage network is AC-coupled, the common-emitter device can be biased at a reduced current level, in principle proportionally to the current amplification contributed by the inter-stage network. Moreover, the reduced output current in the CE stage allows to reduce the Q_1 size by the passive current gain A_i .

A simplified version of the circuit implementation is reported in Figure 3.8. The circuit consists of a power stage (Q_2), a driver stage (Q_1) and an inter-stage LC matching network.

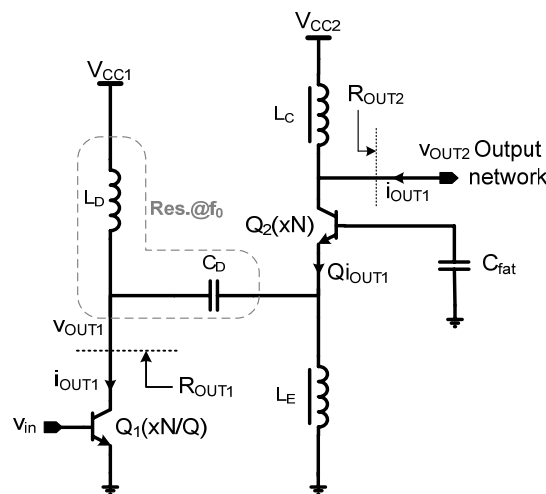


Figure 3.8 Implementation of the AC-Coupled cascode with inter-stage match.

The LC resonator formed by L_D and C_D resonates at the fundamental frequency and, together with Q_2 , provides current amplification. In fact, ignoring the parasitic capacitance at the output of the driver, the RF signal current flowing in L_D and C_D (hence also in Q_2) is equal to the driver's output current multiplied by the loaded quality factor (Q), given by

$$3.16 \quad Q = \left(\frac{1}{Q_D} + \sqrt{\frac{C_D}{L_D}} \frac{1}{g_{m2}} \right)^{-1} \cong \sqrt{\frac{L_D}{C_D}} g_{m2}$$

where g_{m2} is the transconductance of the power transistor and Q_D is the quality factor of inductor L_D . The use of an LC network to perform the current amplification allows for a certain degree of flexibility in the choice of the current gain, the upper bound to the gain being the inductor quality factor.

A large capacitance C_D , hence a low inductance L_D , is desirable in order to minimize the contribution of the inductor quality factor Q_D to the loaded Q and the relative weight of the driver output capacitance compared with C_D , minimizing the sensitivity to parasitic elements. In this way most of the power supplied by the driver stage is delivered to the power stage and the loaded Q is less dependent on process variations. On the other hand, as C_D is increased, current gain decreases. The choice of C_D also impact the driver efficiency. From this point of view, C_D should be set according to the optimum loading condition:

$$3.17 \quad Q \sqrt{\frac{L_D}{C_D}} = R_{OUT,1} \cong \frac{(V_{CC1} - V_{SAT}) A_i}{I_{RF}}$$

where V_{CC1} is the driver voltage supply, A_i is the current gain and I_{RF} the power stage maximum RF current.

Because of the relatively low impedance seen at the driver's output, the collector voltage swing of Q_1 is hence low. As already mentioned, the driver's supply voltage can be reduced in order to maximize its voltage efficiency. Moreover, because of the passive current amplification, the Q_1 size can be scaled by a factor of Q compared to Q_2 . This lowers the contribution of the driver stage to the overall efficiency. As an example we can suppose to have for the output stage the same conditions derived in the previous paragraph. Thus referring to the CB stage, for an output saturation power $P_{SAT} = 0.25W$ and a supply voltage $V_{CC2} = 4.5V$, its output efficiency is $\eta_2 = 61\%$ and the output current is $I_{OUT2} = 143mA$.

Now, if the driver stage is supposed to be biased at $V_{CC1} = 2V$, it is possible to define a supply voltage scaling factor:

$$3.18 \quad \alpha = \frac{V_{CC2}}{V_{CC1}} = 1.8$$

If the interstage network quality factor is supposed to be $Q = 10$ and the Q_1 size is scaled from Q_2 of the same amount, and if the two stages are both biased in

class B, the Q_1 DC current will also scaled from that of Q_2 : $I_{DC1} = I_{DC2}/Q$. As a consequence the DC power dissipated from Q_1 is:

$$P_{DC1} = V_{CC1} I_{DC1} = \frac{V_{CC2}}{\alpha} \frac{I_{DC2}}{Q} = \frac{P_{DC2}}{Q} \quad 3.19$$

In conclusion, the overall drain efficiency is:

$$\eta_{TOT} = \frac{P_{SAT}}{P_{DC1} + P_{DC2}} = \eta_2 \left(\frac{\alpha Q}{1 + \alpha Q} \right) = 57.8\% \quad 3.20$$

This result shows that the passive amplification effect combined to the supply voltage reduction reduces the overall impact of the driver stage, making this solution very attractive for integration.

3.3 Class AB Common Base Design Example.

This paragraph will show the design flow of a Class AB PA in a common base topology with a common emitter driver stage and an inter-stage matching network. The design flow will go through the following steps:

- Common Base output stage;
- Inter-stage matching network;
- Common Emitter driver stage;
- Bias network;
- Design details;
- Layout;
- Test Board;
- Measurements.

The first three steps will consider a single-end solution for simplicity of discussion. Before going through the details of the bias network, the pseudo-differential solution will be considered as the final topology used.

The target of this design is to implement a power amplifier able to deliver 30dBm of maximum linear power (P_{1dB}) with maximum linearity and efficiency. The technology considered is an ST BiCMOS 0.25 μ m with high voltage bipolar transistors. The supply voltage used is 4.5V.

3.3.1 Common Base Output Stage

The following discussion refers to the CB PA stage shown in Figure 3.8 at 3.2.4. That notwithstanding, the design of this stage follows a design flow which is independent from the particular topology used. Once the supply voltage and P_{1dB} are chosen, four main elements must be designed:

- Transistor's bias and size;

- Collector inductor value (to resonate the output capacitance);
- Output optimum resistance.

The design flow is presented schematically in Figure 3.9.

First, a unit transistor's size is chosen, no matter how small or large it is. This because this unity PA can be scaled in order to achieve the desired output power without losing efficiency, once the other parameters are optimized. After that a proper collector inductance value must be chosen to resonate the transistor's output capacitance. This allows to maximize the efficiency, because no power is wasted on the output capacitance. Then a proper DC operating point is selected, because it sets the PA working class and linearity. This bias can be rearranged later, after load resistance optimization (which is intended to maximize the efficiency): since the transistor is capable of delivering a maximum amount of current I_{MAX} the output resistance value must be chosen in order to maximize the voltage efficiency (and so the overall efficiency). This design flow needs some iteration to find the optimum value for each parameter.

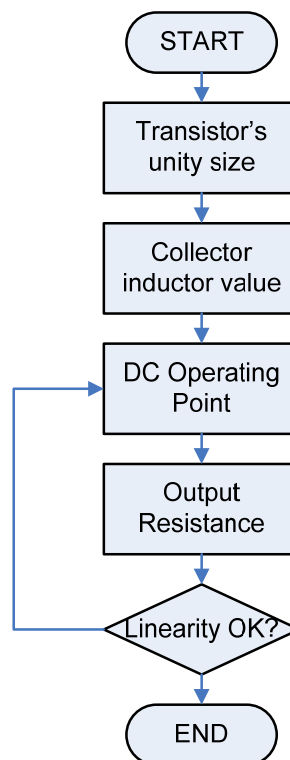


Figure 3.9 PA design flowchart

The first step is the choice of the bias voltage. The DC operating point influences the efficiency and linearity performance, since it controls the conduction angle. Its influence to the efficiency is relative just to the current efficiency, since the voltage efficiency depends mainly on the load resistance. The linearity performance depends by the bias voltage due to the I_c vs. V_{be} characteristic, since the harmonics generation in the collector current is influenced by its nonlinear shape.

Since a Class AB amplification is wanted it is necessary to look at the I_c vs. V_{be} characteristic, in order to choose a right base bias voltage. This characteristic is shown for a unity transistor in Figure 3.10.

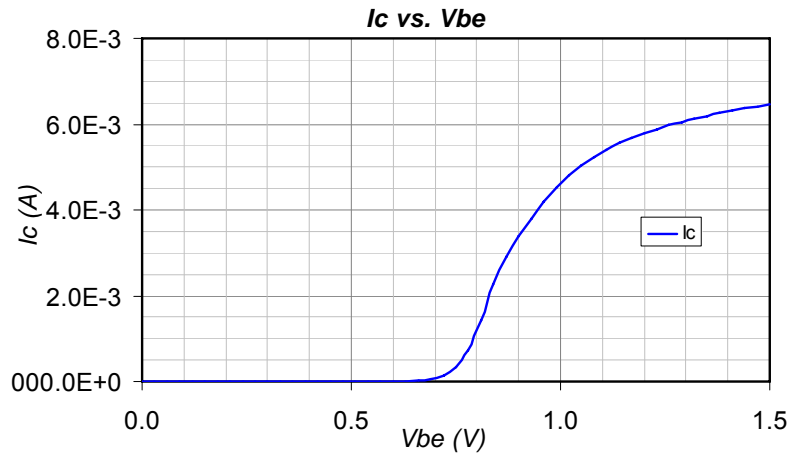


Figure 3.10 Collector Current vs. V_{BE} voltage for a unity transistor

This chart has a shape significantly different from the ideal linear piecewise current shape assumed in Chapter 2. In the real case it is not trivial to find the right bias point. This means that the choice of the optimum bias point for class AB operation has to be performed iteratively, starting from a bias point just above the threshold voltage. The optimum bias will be the one which makes the power gain more linear. In this case, referring to Figure 3.10, the starting point for this iteration has been chosen at $V_{BE} = 0.75V$.

Once the starting bias point is chosen, the next choice is the unity transistor. This selection is quite simple, because no constraint must be satisfied at this point. The only little foresight to apply is that a very small transistor needs a very large collector inductance to resonate the output capacitance. In this case a unity transistor with emitter length $l_e = 5\mu m$ and width $w_e = 0.4\mu m$ and multiplicity 10 has been chosen. With this size a 62nH collector inductor is needed to completely resonate the transistor's output capacitance.

Then, the choice of the output resistance has to be made. This can be performed by sweeping the load resistance until the maximum efficiency is reached. Figure 3.11 shows the maximum efficiency variation for different values of R_{load} . The maximum efficiency is reached with a 750 Ω load. This is the output resistance which maximizes the voltage efficiency. However, small changes in the bias point could be made in order to maximize the linearity. In this case the base bias voltage has been changed from 750mV to 740mV in order to improve linearity. A summary of the performance achieved by this unity amplifier is reported in Figure 3.12 and Figure 3.13.

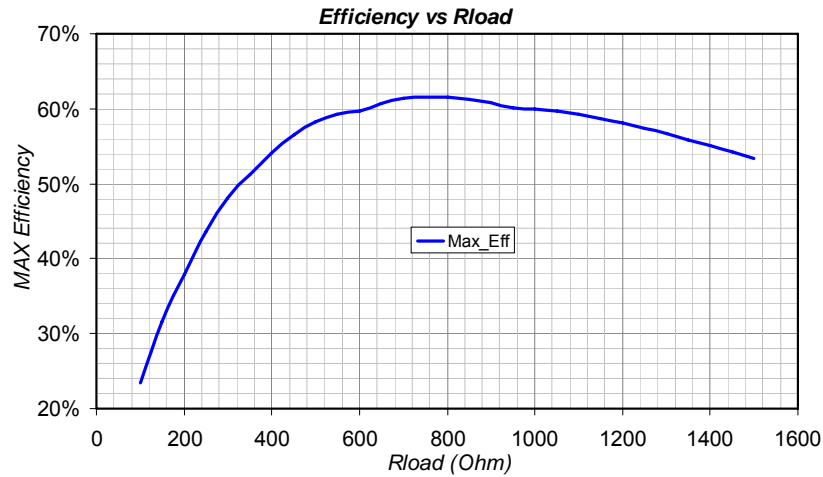


Figure 3.11 Maximum Efficiency for different load resistance values

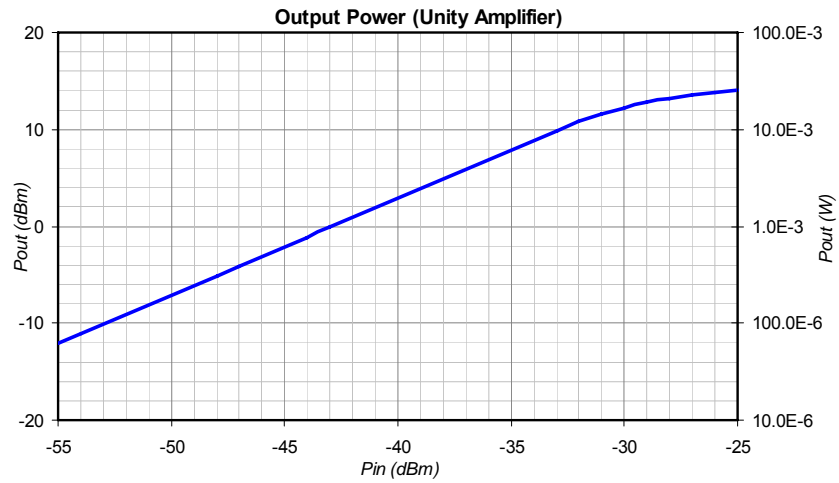


Figure 3.12 Output power of the unity amplifier

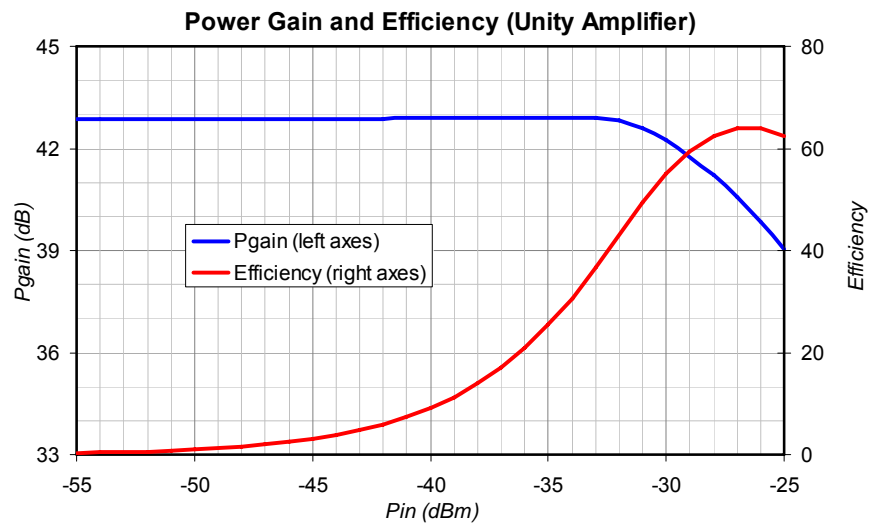


Figure 3.13 Power gain and efficiency of the unity amplifier

At this point the unity amplifier has been optimized. It must be noted that no considerations about the output power have been made so far. This is because it is possible to scale the transistor's size and its output load in order to reach the

desired power level. In this case, the maximum linear power for this amplifier is $P_{1dB_Unity}=19\text{mW}$. If a $P_{1dB} = 0.5\text{W}$ is wanted, the transistor's size must be 50 times larger (the amount P_{1dB}/P_{1dB_Unity}) while the load resistance must be 50 times smaller. This keeps the efficiency and linearity performance close to those achieved in the optimized unity amplifier.

For this particular case the transistor will consist of 500 elements with emitter length $l_e = 5\mu\text{m}$ and width $w_e = 0.4\mu\text{m}$. Its output resistance will be $R_L=30\Omega$ with a 1.3nH collector inductor which resonates the transistor's output capacitance. Figure 3.14 and Figure 3.15 report the performances of the final amplifier.

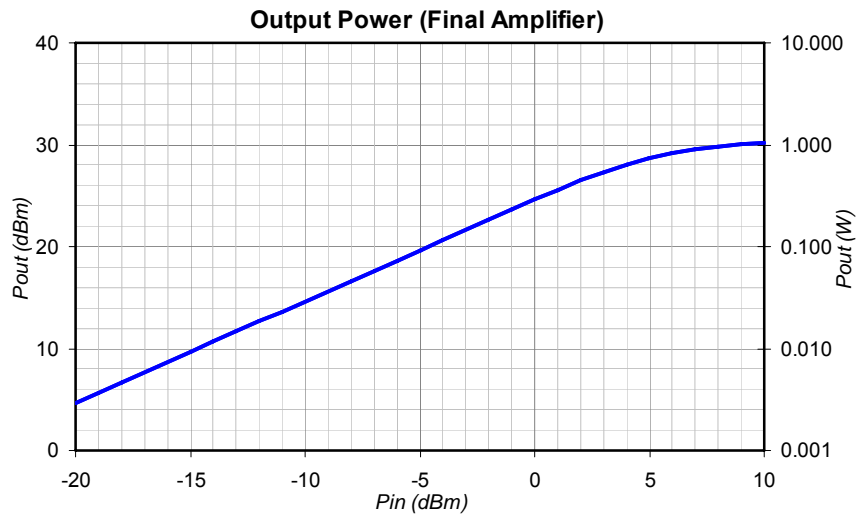


Figure 3.14 Output power of the final amplifier

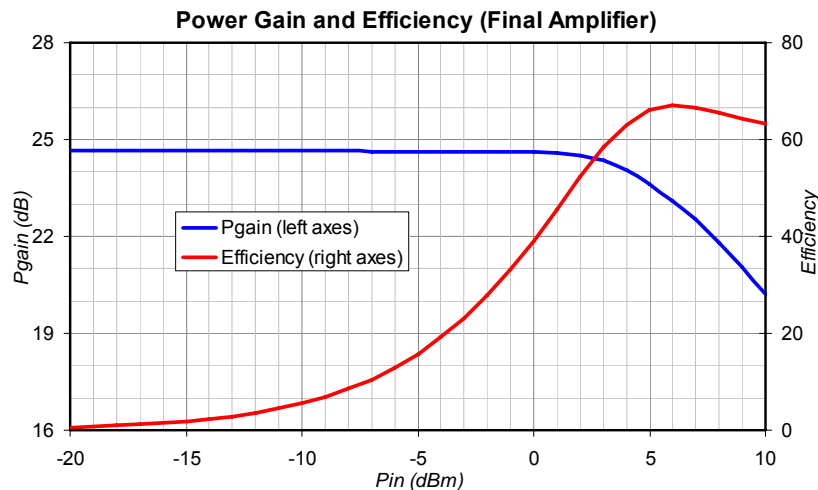


Figure 3.15 Power gain and efficiency of the final amplifier

Looking at the previous graphs it is possible to see that the linearity and efficiency performance are about the same of the unity amplifier but with a maximum linear output power around 28dBm . Following this design flow it is then possible to set the transistor's and the other elements' size starting from a optimized unity amplifier with a few iterations.

3.3.2 Inter-stage Matching Network

The aim of the inter-stage matching network is to provide a passive current amplification from the driver to the output stage. This current gain allows to reduce the driver stage size, reducing its effect to the efficiency reduction. However, a very high current gain is not desirable, since it affects the selectivity (making the amplifier narrowband). Thus, a good trade-off between gain and selectivity must be found.

The output stage can be modelled, as a first approximation, as a resistance equal to $1/g_{m2}$ in series to the C_D capacitance (Figure 3.16), representing the impedance seen at the emitter of the output transistor.

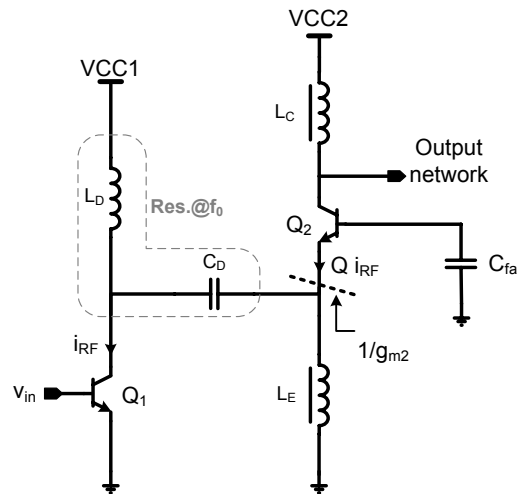


Figure 3.16 Common Base Amplifier with inter-stage matching network

As already mentioned in Paragraph 3.2.4, the passive current amplification is equal to the quality factor of this network:

$$3.21 \quad Q = \frac{1}{\frac{1}{Q_D} + \sqrt{\frac{C_D}{L_D}} \frac{1}{g_{m2}}} \cong \sqrt{\frac{L_D}{C_D}} g_{m2} \quad \left(\text{when } Q_D \gg \sqrt{\frac{L_D}{C_D}} g_{m2} \right)$$

This equation shows that the maximum achievable passive current amplification is limited in practice by the inductor losses (Q_D is the inductor quality factor).

In this design a $Q = 10$ has been chosen. After choosing the current amplification, L_D and C_D must be chosen. Since the inductor quality factor impacts the overall Q , the inductor with the higher Q_D available for the given technology should be chosen. Since C_D and L_D are subjected to both 3.21 and $\omega_0^2 = 1/L_D C_D$, they will fulfil the following relations:

$$3.22 \quad L_D = \frac{Q}{\omega_0 g_{m2}} \left(1 - \frac{Q}{Q_D} \right)$$

$$C_D = \frac{1}{(2\pi f_0)^2 L_D} \quad 3.23$$

Table 3.1 shows the L_D and C_D values for several Q , supposing to have $g_{m2}=3.5S$, $f_0=1.95GHz$, and $Q_D=15$:

| Q | L_D (H) | C_D (F) |
|----|-----------|-----------|
| 1 | 25.0E-12 | 266.6E-12 |
| 2 | 53.8E-12 | 123.8E-12 |
| 3 | 87.4E-12 | 76.2E-12 |
| 4 | 127.2E-12 | 52.4E-12 |
| 5 | 174.9E-12 | 38.1E-12 |
| 6 | 233.2E-12 | 28.6E-12 |
| 7 | 306.1E-12 | 21.8E-12 |
| 8 | 399.8E-12 | 16.7E-12 |
| 9 | 524.7E-12 | 12.7E-12 |
| 10 | 699.6E-12 | 9.5E-12 |
| 11 | 961.9E-12 | 6.9E-12 |
| 12 | 1.4E-9 | 4.8E-12 |
| 13 | 2.3E-9 | 2.9E-12 |
| 14 | 4.9E-9 | 1.4E-12 |

Table 3.1

The previous table shows that a $L_D = 700pH$ and $C_D = 9.5pF$ are needed to obtain the desired passive current amplification. It should be note that C_D is the sum of the capacitor placed into the circuit and the output capacitance of the driver stage.

Another constraint must be taken into account. In fact, in order to optimize the driver stage, its output resistance must fulfil the following relation:

$$R_{DRV,opt} = \frac{V_{CC} - V_{SAT}}{\frac{I_{RF}}{Q}} = \sqrt{\frac{L_D}{C_D}} \frac{1}{Q_D^{-1} + \sqrt{\frac{C_D}{L_D}} \frac{1}{g_{m2}}} \quad 3.24$$

This relation shows that the driver supply voltage is related to its output impedance for an optimum design. This means that an higher output resistance reflects to an higher supply voltage making the driver stage less efficient. The element which mainly affect this resistance is the inductor quality factor: the lower it is, the higher $R_{DRV,opt}$ will be (for the same current amplification amount). Table 3.2 shows a comparison between two different Q_D values.

| $Q_D=15$ | | | | $Q_D=25$ | | | |
|----------|-----------|-----------|----------------------------|----------|-----------|-----------|----------------------------|
| Q | L_D (H) | C_D (F) | $R_{DRV,opt}$ (Ω) | Q | L_D (H) | C_D (F) | $R_{DRV,opt}$ (Ω) |
| 2 | 53.8E-12 | 123.8E-12 | 1.3 | 2 | 50.7E-12 | 131.4E-12 | 1.2 |
| 3 | 87.4E-12 | 76.2E-12 | 3.2 | 3 | 79.5E-12 | 83.8E-12 | 2.9 |
| 4 | 127.2E-12 | 52.4E-12 | 6.2 | 4 | 111.0E-12 | 60.0E-12 | 5.4 |
| 5 | 174.9E-12 | 38.1E-12 | 10.7 | 5 | 145.7E-12 | 45.7E-12 | 8.9 |
| 6 | 233.2E-12 | 28.6E-12 | 17.1 | 6 | 184.1E-12 | 36.2E-12 | 13.5 |
| 7 | 306.1E-12 | 21.8E-12 | 26.3 | 7 | 226.7E-12 | 29.4E-12 | 19.4 |
| 8 | 399.8E-12 | 16.7E-12 | 39.2 | 8 | 274.3E-12 | 24.3E-12 | 26.9 |
| 9 | 524.7E-12 | 12.7E-12 | 57.9 | 9 | 327.9E-12 | 20.3E-12 | 36.2 |
| 10 | 699.6E-12 | 9.5E-12 | 85.7 | 10 | 388.7E-12 | 17.1E-12 | 47.6 |
| 11 | 961.9E-12 | 6.9E-12 | 129.6 | 11 | 458.1E-12 | 14.5E-12 | 61.7 |
| 12 | 1.4E-9 | 4.8E-12 | 205.7 | 12 | 538.1E-12 | 12.4E-12 | 79.1 |
| 13 | 2.3E-9 | 2.9E-12 | 362.1 | 13 | 631.6E-12 | 10.5E-12 | 100.6 |
| 14 | 4.9E-9 | 1.4E-12 | 840.0 | 14 | 742.0E-12 | 9.0E-12 | 127.3 |

Table 3.2

The previous table shows that for a current amplification of 10, the driver output resistance changes significantly from $Q_D=15$ to $Q_D=25$ and also the L_D and C_D value. This is a big issue for technologies where inductors with high quality factor are not available. It is also possible to decrease the value of Q, but this solution reduces the power added efficiency as already mentioned. Thus, a good trade-off between Q, Q_D and L_D must be found.

The efficiency reduction due to the effect of Q_D is reported in Figure 3.17. Here the efficiency of the common base stage is compared to the efficiency including the driver stage for different Q_D values.

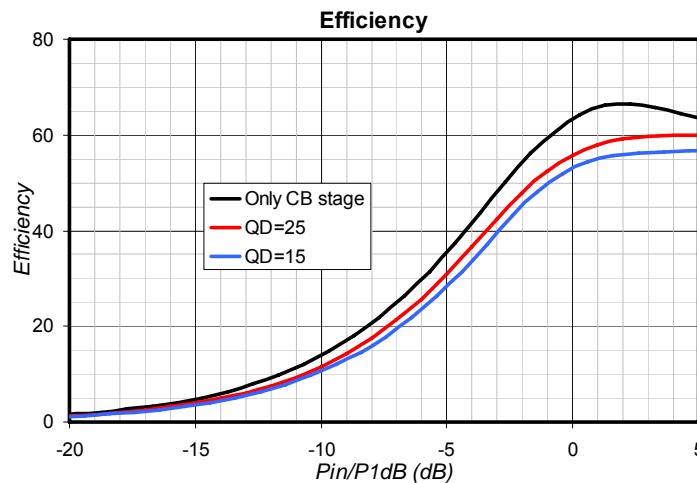


Figure 3.17 Efficiency reduction due to the inductor quality factor

The efficiency reduction due to the driver stage has been already explained in Paragraph 3.2.4. The efficiency reduction due to decreasing reduction of Q_D is not dramatic. The bias voltage is different between the two cases because the optimum V_{DD} follows the 3.24. Since $V_{CC}-V_{SAT}$ is the voltage swing at the driver's collector it is necessary to give the right headroom to the driver stage.

Figure 3.18 shows the voltage swing at the driver's collector for the two cases. The maximum linear voltage swing (at $P_{in}/P_{1dB}=0$ dB) is equal to 2.8 V for $Q_D=15$ and 1.6V for $Q_D=25$. Thus a $V_{CC}=3.8V$ and 2.6V is needed respectively in order to have an optimum design for the driver stage.

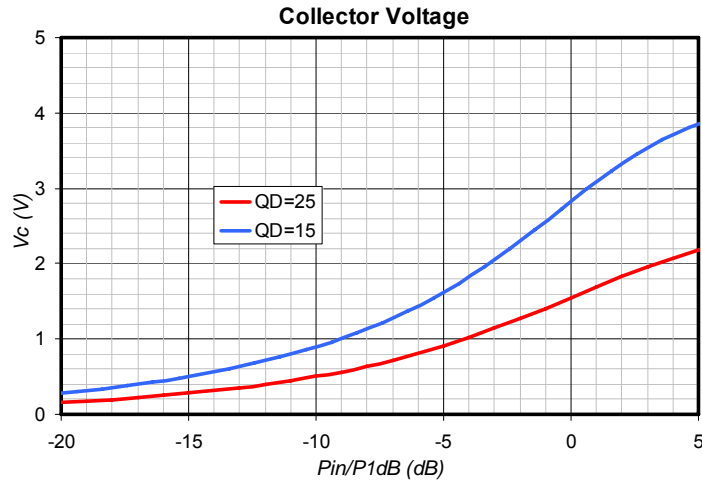


Figure 3.18 Dependency of the collector voltage from the inductor quality factor

3.3.3 Common Emitter Driver Stage

The driver stage design is straightforward once the proper current amplification factor has been chosen.

Since a $Q=10$ has been chosen, the driver's size could be 10 times smaller than the output stage. This would mean that the driver and the output stage will have the same behavior in terms of power compression (if an optimum driver stage has been designed). However, because of the issues related to the inter-stage matching network, an optimum driver output impedance it is not guaranteed because of the variability on the capacitor and inductor values and quality factor. These could lead the driver stage to work in under-optimum conditions, even to an anticipated power compression. In order to avoid this scenario, the driver stage should be able to deliver the right RF current to the output stage until this one reaches its 1-dB compression point. This means that gain compression on the driver stage must take place at a power level higher than the maximum power requested from the output stage. This guarantees that the total gain compression is due only to the output stage.

For these considerations the driver power compression has been chosen to be at least 1dB further than the output stage one's. Assuming that the output current (and so the maximum power) of the driver stage scales with the transistor size, a difference of 1dB in power reflects into a transistor 1.2 times bigger than the optimum size. Since $Q=10$ if the driver stage is set to be 10 times smaller than the output stage than power compression for both the stages happens at the same input power.

Figure 3.19 shows the driver and output stage power gain referred to the latter's 1dB compression point (0dB on the x-axes). The driver stage (red curve) has a compression point 1dB further than the output stage, as expected.

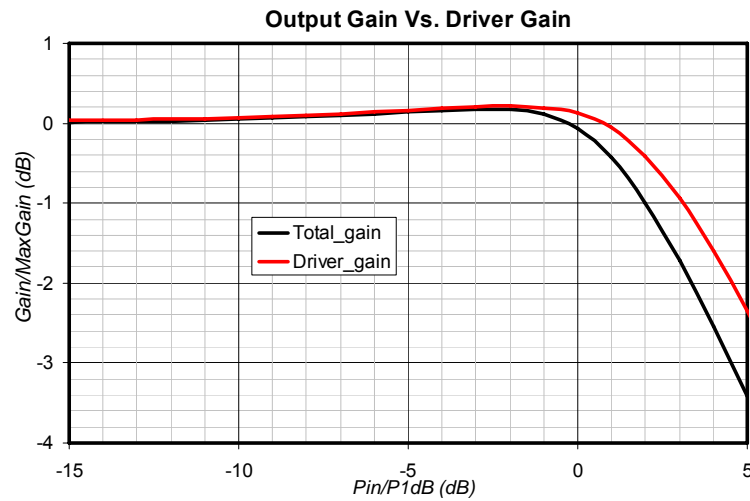


Figure 3.19 Power gain of the driver and the output stage

Since the driver stage efficiency has a relatively low impact on the overall efficiency, the driver stage can be biased in a more linear class, with benefits in terms of linearity performance of the overall amplifier.

While for the output stage stability is not an issue (because of the common-base configuration), this must be taken into account for the driver. This issue is prevented using a cascode topology: this reduces the driver efficiency, but has a limited impact on the overall efficiency. Moreover, because of the high supply voltage of the driver stage (due to its relatively high output impedance, as shown in Paragraph 3.3.2), a cascode solution can prevent the over-voltage issues at the driver's collector.

The right voltage to bias the cascode base is equal to $2V_{SAT}$. In fact, this would be the minimum reachable collector voltage before entering the saturation region under optimum conditions. An higher base voltage will increase the minimum voltage floor, decreasing the maximum swing achievable, while a lower base voltage will keep the underneath transistor near to the triode region, reducing the maximum linear power.

Figure 3.20 shows the impact on the efficiency due to the cascoded driver stage. It is possible to see that the efficiency reduction is acceptable, since at the one dB compression point it changes from 54% to 51%.

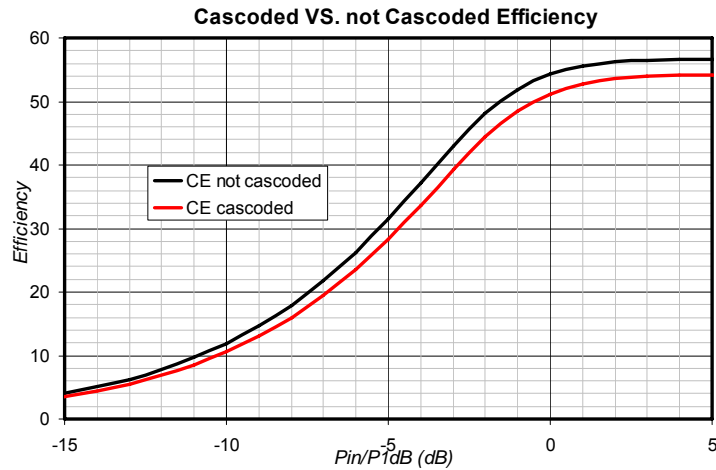


Figure 3.20 Efficiency reduction due to the cascoded driver stage

3.3.4 Bias Network

The bias network must fulfil two aims:

- Provide a low base impedance at DC and all of the harmonics of the signal;
- Supply the DC base current (which increases with the input power at the higher levels);

The DC operating point can be generated in two ways: using a voltage source or a current mirror. The first solution is able to provide a low base impedance and it has a good compliance to the base current requested, but it is not the best solution. Since a precise voltage must be provided in order to correctly select the amplification class, it is difficult to generate a bias voltage which can follow the quiescent current variation due to temperature changes. In fact, looking at Figure 3.21, it is possible to see how the DC collector current changes with temperature. Thus, if the operating point is set by a base voltage, the amplification class will change with temperature, affecting the linearity and efficiency performance. A current mirror solution is then the best candidate, and the topology chosen is shown in Figure 3.22.

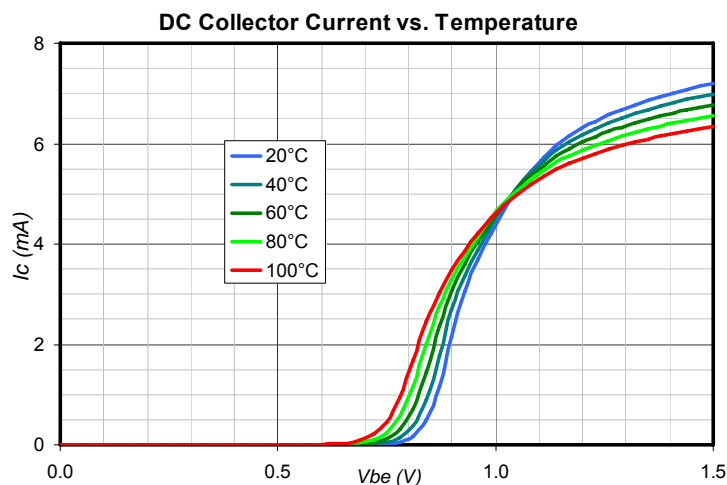


Figure 3.21 Effect of temperature on the DC collector current

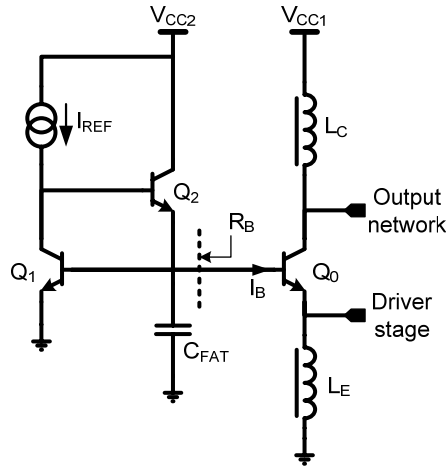


Figure 3.22 Bias stage implementation

The current mirror is formed by Q_1 and Q_0 (which is the power transistor) while the purpose of Q_2 is double: to supply the higher base current requested at the higher levels and provide a low base impedance. In fact the impedance seen by Q_0 at its base is:

3.25

$$R_B = \frac{1}{\frac{1}{r_{\pi_1}} + g_{m_2}(1 + g_{m_1}r_{o_1})} \approx \frac{1}{g_{m_1}g_{m_2}r_{o_1}}$$

where g_{m_1} and g_{m_2} are the transconductance gains of Q_1 and Q_2 , r_{o_1} and r_{π_1} are respectively the base-emitter and the output resistance of Q_1 . This gives a low impedance at DC, while at the higher frequencies it is synthesized by C_{FAT} , which is a very large integrated capacitor. The value of this capacitor should be very high, since a 50pF capacitor has an impedance of 1.6Ω at 2GHz.

The base bias voltage of the cascode transistor in the driver stage can be supplied by a simple voltage source, since there is not the need to provide a very precise base voltage because the amplification class is set by the common emitter transistor. This one can be biased by a current mirror in order to be less sensitive to temperature variations. In this case a low base impedance is not needed because the voltage at the collector has a low variation thanks to the cascode action. Anyway the same topology of Figure 3.22 has been chosen for simplicity.

3.3.5 Design Details

The complete pseudo-differential schematic is reported in Figure 3.23.

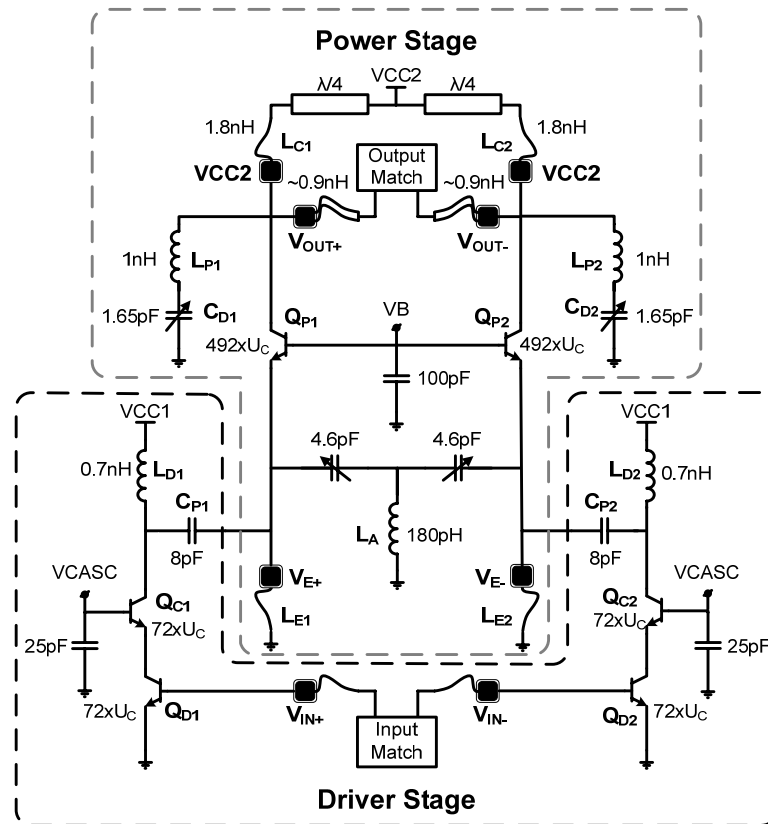


Figure 3.23 Complete pseudo-differential schematic

This solution has some details that were not discussed in the previous paragraph. Here they are shown referring to the driver and the power stage parts of the circuit.

3.3.5.1 Power stage

In this stage the “choke” inductors (L_C and L_E in Figure 3.23)

The $2f_0$ resonant network at the output has been realized with a series LC network (L_D and C_D); the capacitor is made by a selectable array of capacitors and provides a fine tuning around the second harmonic. This network is not critical, since it must give an impedance at $2f_0$ which is low compared to the impedance seen at the same point (which is the output impedance). A different remark concerns the $2f_0$ resonant network at the output transistor’s emitter. Here the impedance synthesized must be very low compared to the $1/g_{m2}$ impedance. This involves the use of an inductor with a very high quality factor.

The $2f_0$ resonant network at the output has been realized with a series LC network (L_D and C_D); the capacitor is made by a selectable array of capacitors and provides a fine tuning around the second harmonic. This

network is not critical, since it must give an impedance at $2f_0$ which is low compared to the impedance seen at the same point (which is the output impedance). A different remark concerns the $2f_0$ resonant network at the output transistor's emitter. Here the impedance synthesized must be very low compared to the $1/g_{m2}$ impedance. This involves the use of an inductor with a very small value and an high quality factor.

Since the second harmonic is a common mode signal, the $2f_0$ resonant network can be realized in the way depicted in Figure 3.24.

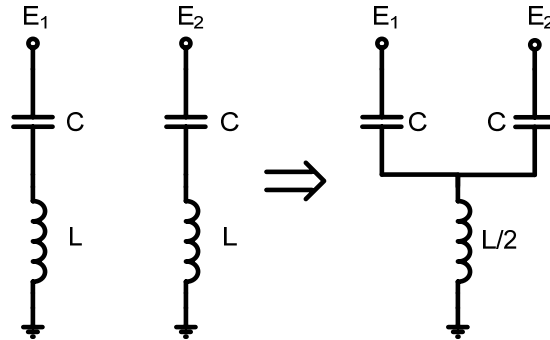


Figure 3.24 Implementation of the 2nd harmonic low impedance

This allows the use of a single small inductor with a very high quality factor. This inductor has been customized in order to reach the desired performance.

For what concerns the power transistors, a 492 unity elements has been used instead of 500 for layout reasons that will be discussed later. The V_B is derived by a current mirror having the topology described in Figure 3.22, with Q_1 and Q_2 made by 4 and 20 unity elements respectively.

3.3.5.2 Driver Stage

In this stage the capacitance C_p is smaller than the 9.5pF obtained with the simulations shown in Table 3.2. This because the output capacitance of the cascode transistor and the residual capacitance at the emitter of the power stage add to this capacitance. Thus a 8pF capacitor is necessary.

The driver transistors consists of 72 unitary elements. The V_{CASC} voltage is set by a voltage source outside the circuit, while the bias point at Q_{D1} and Q_{D2} is set by a current mirror. Since the signal is fed into the base, an high impedance must be provided by the current mirror. This is achieved using an inductor outside the chip, as shown in Figure 3.25.

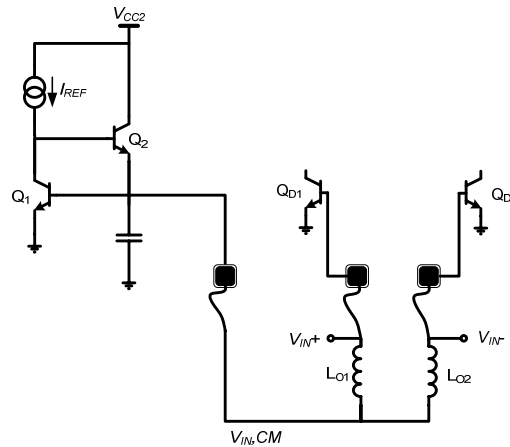


Figure 3.25 Bias of the driver stage

This configuration gives the possibility to measure the base DC voltage of Q_1 and Q_2 . In this way it is also possible to indirectly measure the internal die temperature, as will be shown in the next paragraphs. The input matching network is off-chip, and it will be discussed later.

3.3.6 Layout

The complete layout of the schematic reported in Figure 3.23 is shown in Figure 3.26. The die area is 2.76mm^2 .

The transistors in the driver stage (Q_{D1} , Q_{D2} , Q_{C1} , Q_{C2}) are totally made by $72 \times 4 = 288$ unit transistors ($Q_D + Q_C$ in Figure 3.26). Moreover the 24 elements of the bias stage (4 elements for Q_{DB} and 20 elements for Q_{DF}) add to the former to a total of 312 elements. All these elements have been drawn in a single symmetric structure, in order to minimize the mismatch (especially in the current mirror).

The transistors of the power stage have separate layout (Q_{P1} and Q_{P2}). Since each of them is made by 492 unit element plus 14 elements of the bias stage, they would have a wide area if placed together, creating a significant difference on the temperature between the core and the periphery of the structure. Moreover this topology allows to have limited connections (i.e. less metal layers used) in order to reduce losses due to the interconnections.

The driver and the output stage have separated on-chip ground connections, in order to prevent parasitic feedback between the two stages which can generate oscillations. Moreover a large amount of ground pads for the two stages have been used, in order to limit the parasitic inductance due to the bondwires.

L_A is an anchor shaped customized inductor with high quality factor. It synthesizes an inductance of 180pH.

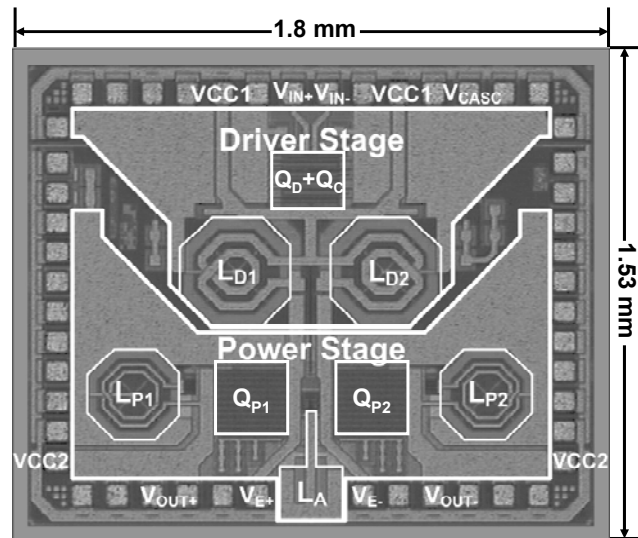


Figure 3.26 Complete layout

3.3.7 Test Board

The test board has three main purposes:

- Power supply filtering;
- Synthesize the input and output matching;
- Provide a thermal ground to the die.

The first aim can be achieved by using filtering capacitors in order to filter out high and low frequency noise in the power supply lines.

For what concern the impedance matching it should be note that the input and the output matching have different impact to the circuit performance.

The input matching network goal is to maximize the power fed into the driver stage. This is achieved by matching the output impedance of the signal source (typically 50Ω) with the input impedance of the matching network. There are two ways to obtain a 50Ω impedance from the input impedance of the driver stage:

1. transforming the input impedance of the driver stage in 50Ω using an LC network;
2. transforming the input impedance of the driver stage in an high impedance which will be in parallel to a real 50Ω resistor soldered on the board.

The first approach maximizes the power fed into the transistors. With the second approach the transistor is treated as a voltage amplifier (the power flows into the soldered resistance), and it is a simpler method to obtain the input matching. In fact the bandwidth of an impedance transformation (i.e. the overall quality factor) is directly dependent on the impedance transformation network. Since the transistor has an high input impedance, it is very difficult to make a wideband impedance transformation. Thus, the second approach (which gives a broadband

matching) has been chosen to synthesize the input impedance. Moreover a poor input matching just reduces the PA power gain, affecting the power added efficiency, but if the power gain is higher than 10dB the PAE doesn't change dramatically.

The output impedance matching has a major impact to the circuit performance. The impedance synthesized by this network determines the output power and strongly affects the efficiency. An output impedance different from the expected prevents the circuit from operating in the optimum conditions, reducing the efficiency.

The output network has been made exploiting the bondwires which connect the output pad to the test board. The single-ended schematic is reported Figure 3.27.

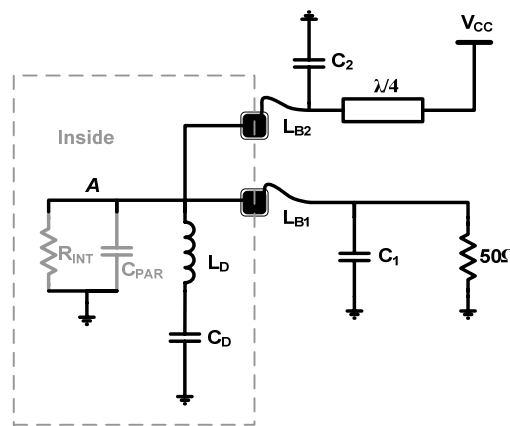


Figure 3.27 Output matching network

Here R_{INT} and C_{PAR} represent the small signal output impedance and output capacitance of the output transistor. In particular, C_{PAR} takes into account also the parasitic capacitance due to the layout. L_D and C_D are the series resonant network at $2f_0$ (already described). The target is to transform the 50Ω antenna impedance into a purely resistive 30Ω at node A . In order to achieve this result, the bondwire L_{B1} and L_{B2} are used. In particular, L_{B2} (with C_2) resonates part of the residual capacitance due to C_{PAR} and the resonant network L_D-C_D , while L_{B1} (combined with C_1) operates the impedance transformation in conjunction with the internal residual capacitance. The partial resonance can be made because the $\lambda/4$ transmission line generates an high impedance, which allows the series connection between L_{B2} and C_2 .

Taking into account that $L_D=1\text{nH}$ and $C_D=1.67\text{pF}$ (which resonate at 3.9GHz) they give a residual capacitance C_{RES} at $f_{RES} = 1.95\text{GHz}$ which is given by:

$$C_{RES} = \frac{C_D}{1 - \omega_{RES}^2 L_D C_D} = 2.23\text{pF} \quad 3.26$$

This capacitance adds to C_{PAR} (which is estimated to be 11pF) giving a total capacitance $C_{TOT}=13.23\text{pF}$. This capacitance can be resonated with a parallel

inductance of 0.5nH. This value can be synthesized by L_{B2} and C_2 . However, a different approach has been used, since a very precise and low inductive value is difficult to achieve on the test board. Thus C_{TOT} has been partially resonated, and the residual capacitance makes a pi-network as shown in Figure 3.28.

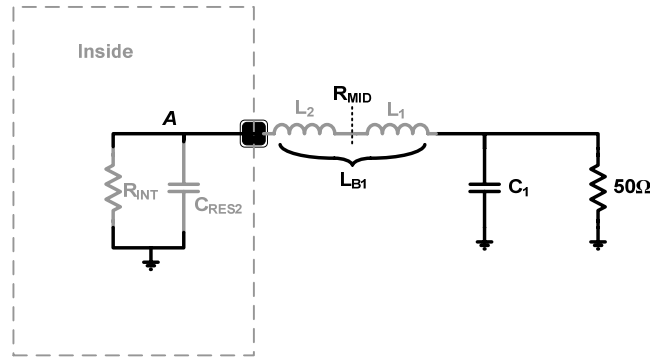


Figure 3.28 Output matching network synthesis

Choosing a C_2 of 20pF, if the series resonance with L_{B2} is wanted to be at 1GHz (in order to keep it far from the fundamental frequency), it must be:

$$3.27 \quad L_{B2} = \frac{1}{(2\pi f_s)^2 C_2} = 1.3nH$$

and the residual inductance at f_{RES} will be:

$$3.28 \quad L_{RES} = \frac{\omega_{RES}^2 L_{B2} C_2 - 1}{\omega_{RES}^2 C_2} = 0.97nH$$

As previously stated, this inductance resonates out part of C_{TOT} , and the residual capacitance C_{RES2} is given by:

$$3.29 \quad C_{RES2} = \frac{1 - \omega_{RES}^2 C_{TOT} L_{RES}}{\omega_{RES}^2 L_{RES}} = 6.36pF$$

The L_{B1} and C_1 values has to be chosen. Referring to Figure 3.28, L_{B1} can be split in two inductances L_1 and L_2 and the intermediate resistance seen between the two is named R_{MID} . Choosing an L_2 value of 1nH (so that L_1 will add to this value to make a feasible L_{B2}), R_{MID} will be:

$$3.30 \quad R_{MID} = \frac{L_2}{R_{INT} C_{RES}} = 5.24\Omega$$

It is now possible to calculate C_1 , L_1 and consequently L_{B1} :

$$3.31 \quad C_1 = \frac{1}{\omega_{RES} R_L} \sqrt{\frac{R_L}{R_{MID}} - 1} = 4.77pF$$

$$L_1 = R_{MD} R_L C_1 = 1.25nH$$

$$L_{B1} = L_1 + L_2 = 2.25nH$$

The simulation performed with Agilent ADS of the impedance synthesized at node A by this matching network is reported in Figure 3.29.

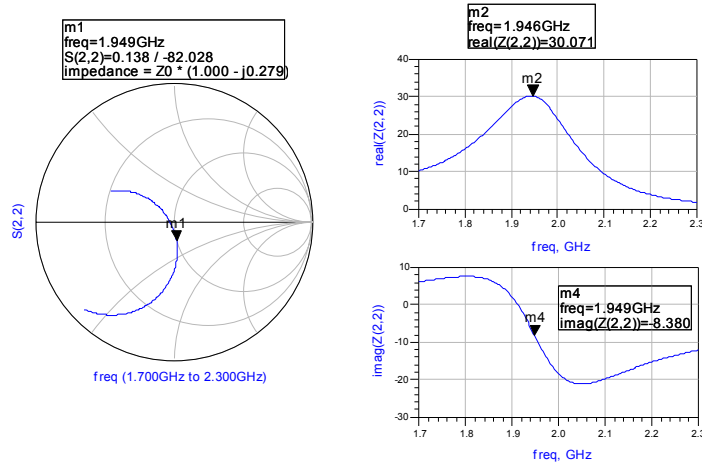


Figure 3.29 Simulations of the output matching network

The impedance seen at the transistor's collector has a small change in the desired band (1.92-1.98GHz) with a small reactive part. However this is the simulation with the previously calculated values: in the real world these can be quite different, since they are synthesized by bondwires and real elements soldered onto the test board. Thus some other simulation has been performed looking at the effect to the impedance transformation due to the variability of L_{B1} , L_{B2} and C_1 and they are reported in Figure 3.30:

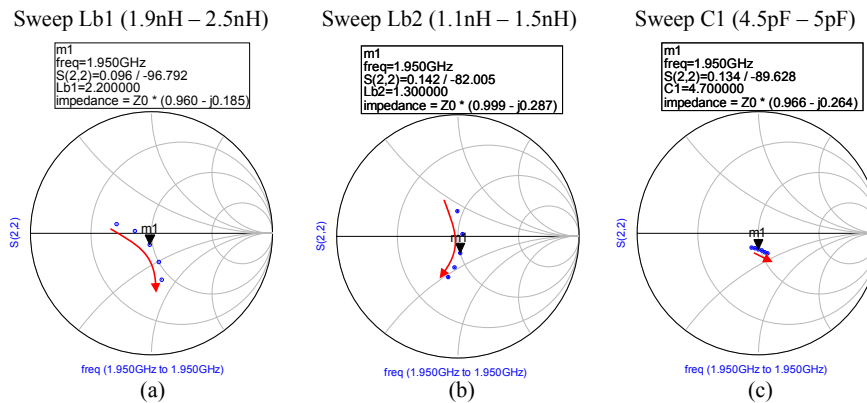


Figure 3.30 Effect of the elements variation

C_1 has a small effect to the impedance (Figure 3.30c) while L_{B1} and L_{B2} have a major influence. However, it is possible to see that if L_{B1} is kept slightly larger than the optimum value and L_{B2} is chosen slightly lower, the impedance remains relatively close to the unitary circle, limiting the negative impact in terms of efficiency.

3.3.8 Simulation Results

Here the simulation of the pseudo-differential circuit are reported. This simulations are performed for a die temperature of 70 Celsius. The transistors are modeled with the HICUM model, which takes into account the effect due to temperature and self-heating, and it is more accurate at high current levels.

Figure 3.31 and Figure 3.32 show the output power and the power added efficiency, respectively. On the X axes the input power is normalized to the 1dB compression point, so that the power and efficiency at 0dB represent the maximum linear power and efficiency (which are 30dBm and 46% respectively).

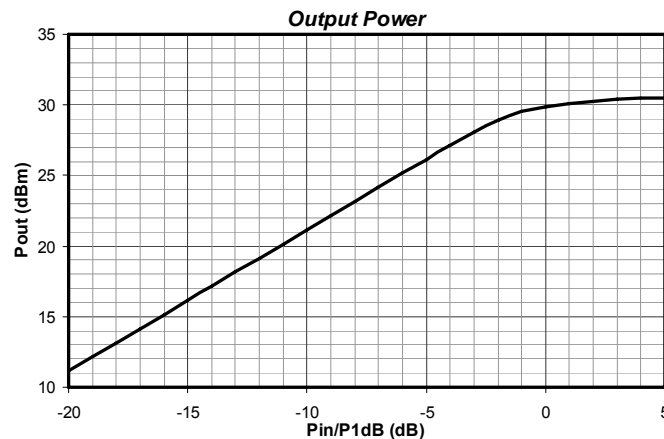


Figure 3.31 Simulated output power of the pseudo-differential schematic

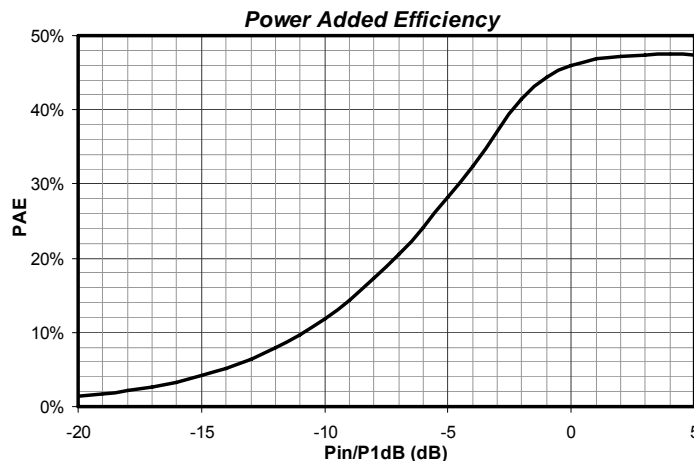


Figure 3.32 Simulated Power Added Efficiency of the pseudo-differential schematic

Looking at the efficiency, it is possible to see that it is slightly lower than the one reported in Figure 3.20 (which was around 50%) at the 1dB compression point. This because now the bias networks and the associated power consumption of the driver and the output stages are taken into account.

Figure 3.33 and Figure 3.34 show the power gain and the DC current dissipated by the driver and the output stage. The power gain looks flat showing a good linearity. Looking at the curves of the DC current it is possible to see that the DC current of the output stage shows a saturation, while the driver stage is still in its

linear region. This is because the driver stage works in a near class-A condition, and the collector current is not clipped due to overdrive conditions as in the class AB output stage.

These simulation results will be compared to the measurements in the next paragraphs.

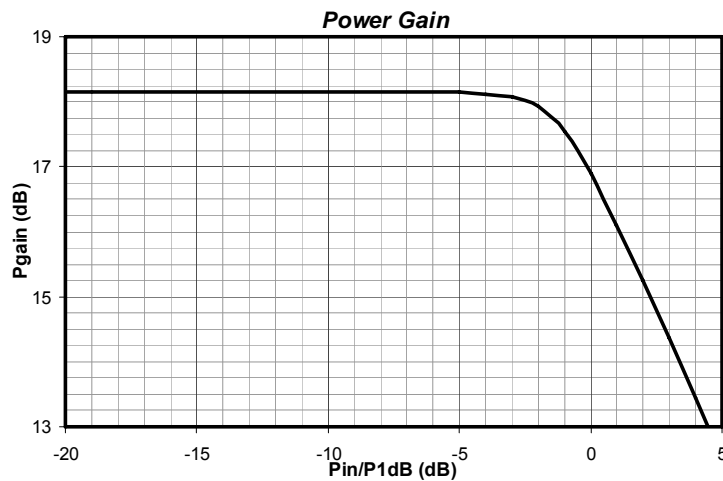


Figure 3.33 Simulated Power Gain of the pseudo-differential schematic

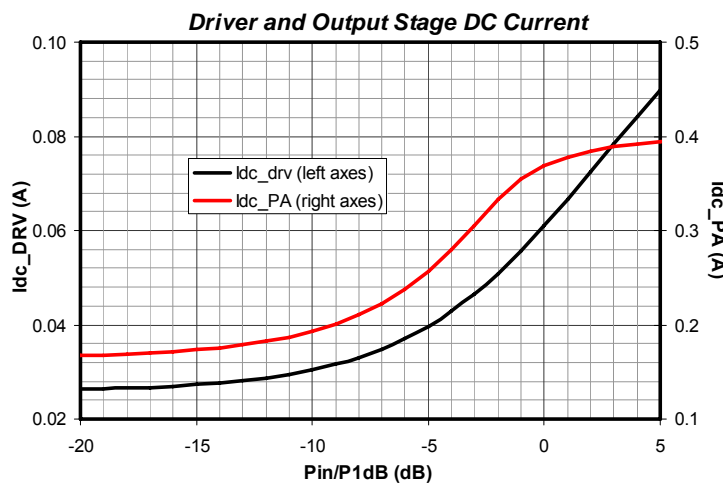


Figure 3.34 Simulated DC currents of the pseudo-differential schematic

3.3.9 Measurement Setup

Figure 3.35 shows the measurement setup used. The amplifier is biased with two different DC voltage sources which provide V_{CC1} and V_{CC2} for the driver and the output stage.

The input signal is fed to the circuit using a hybrid coupler, which converts the signal generator output to a differential signal.

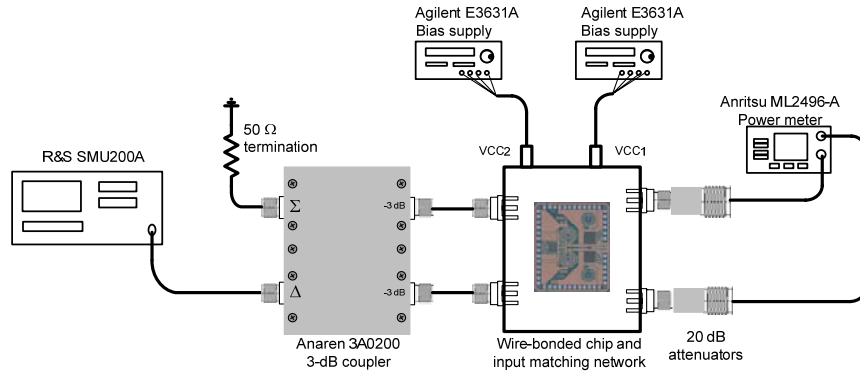


Figure 3.35 Measurement setup

The output signal is measured by a power meter which measures the two differential outputs separately. The outputs are attenuated by a 20dB attenuator which limits the signal at the power meter input in order to prevent damages.

3.3.10 Measurement Results

Here the measurements of output power and gain (Figure 3.37 and Figure 3.36), efficiency (Figure 3.39) and dissipated current (Figure 3.38) at 1.8GHz are reported to the input power normalized to the 1dB compression point.

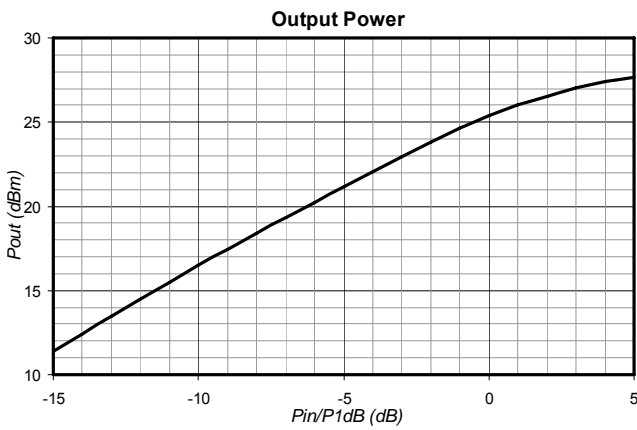


Figure 3.37 Measured output power

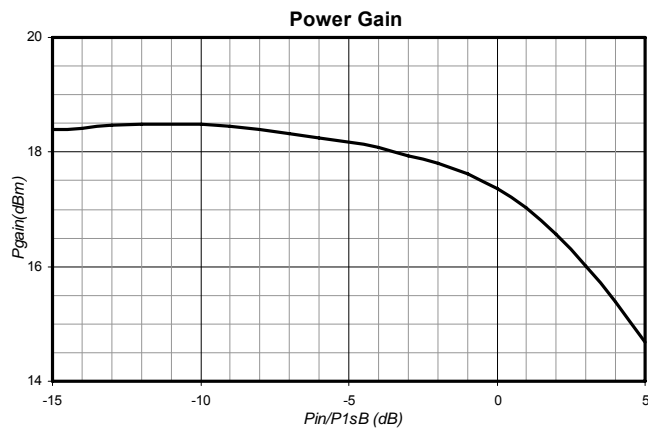


Figure 3.36 Measured Power Gain

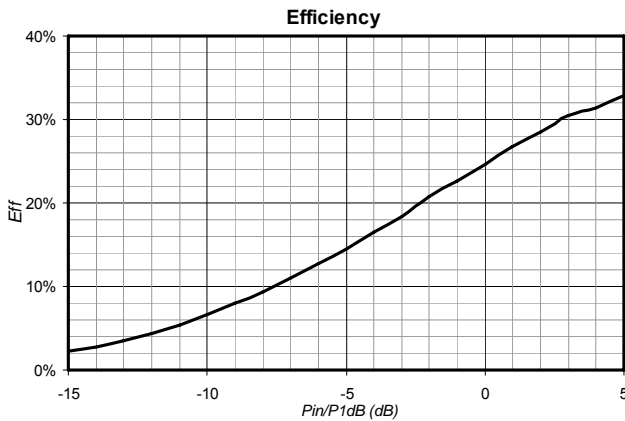


Figure 3.39 Measured Power Added Efficiency

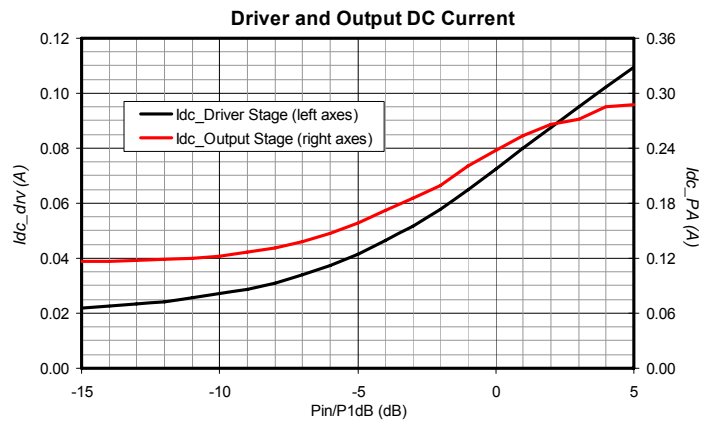


Figure 3.38 Measured DC current

These measurements have been performed with a $V_{CC1}=V_{CC2}=4.5V$. The measured output power at the 1dB compression point (25.5dBm) is very low compared to the simulated output power (30dBm) and also the power added efficiency (25% vs. 46%) which is strictly related to the output power. Moreover the DC current of the output stage (240mA) is well below compared to the simulated one (370mA). Finally, the power gain is not flat, showing early power compression.

Several effects can explain the reduced output power. The main causes that were identified are:

- Lower inter-stage current gain
If the current gain between the driver and the output stage is too low, the output current is reduced limiting the output power.
- Reduced load resistance
A reduced load resistance, for a given signal current, leads to a reduced output voltage swing and lower output power. This leads to a reduced voltage (and power) efficiency.
- Thermal effects
As will be shown in more detail in the Appendix, thermal effects induce a reduced power.

None of the above effects alone is sufficient to explain the observed reduction in output power.

A lower inter-stage current gain is supported by the measurements results. In fact this current gain sets the output current flowing into the load. With a reduced RF current also the DC current of the power stage will be lower than expected. Recalling the basis of an ideal class B amplifier, the ratio between the RF and the DC current is fixed ($\pi/2$). Since a class AB amplifier works like a class B when it is near to the compression point, we can expect a lower output current if the quiescent current is low. However this effect alone determines a reduction in output power of 1.5dB, which is much less than what is measured.

A reduced load resistance, for a given signal current, leads to a reduced output voltage swing. The presence of reduced load can be revealed by changing the supply voltage of the power stage. If a reduction on the supply voltage doesn't change the saturation power, this means that the voltage swing at saturation is low due to the low load resistance. This has been actually revealed during the measurements, where the supply voltage of the output stage has been reduced until the saturation power changed and we have found this limit at 3V. More accurate simulation have been performed including the EM simulation of the test board and real models of the components placed onto the board. These simulations actually revealed a lower output load synthesized by the matching network, as is possible to note looking at the next figures.

Output Resistance

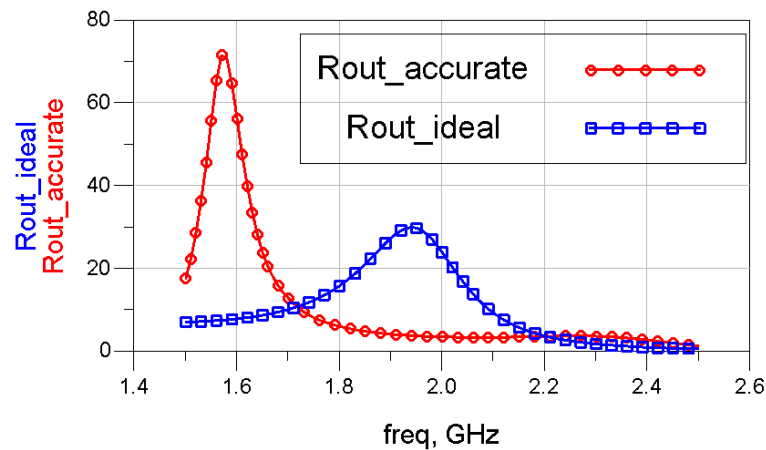


Figure 3.40 Effect of the test board on the output resistance

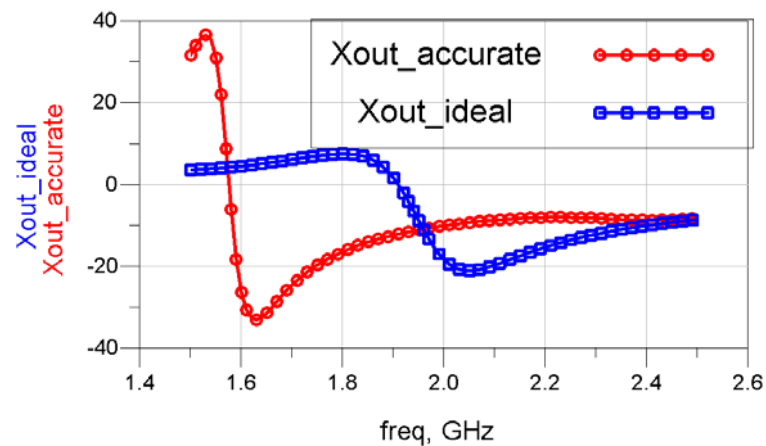


Figure 3.41 Effect of the test board on the output reactance

Thermal effects due to high thermal resistance of the testing board cause a reduction of the RF current which leads to a reduced output power and nonlinear effects, as explained in Appendix. The die temperature has been measured at full output power using a thermo-graphic camera that revealed a temperature of around 200°C at the output transistors. This is a much more higher temperature compared to that used in simulations (70°C): this means that the testing board shows an high thermal resistance.

All of these effects have been included in a simulation test bench which allows to predict more accurately the real circuit behavior. The results are shown in the next figures, referring to a supply voltage for the power stage of 3V, and they are compared to the simulation of a test board with lower thermal resistance.

The measurements show actually some more output power and efficiency compared to the simulated one considering a 180°C/W thermal resistance. This is due to the fact that the driver and the output stage are simulated at the same temperature, while in reality they have different temperatures. In fact, even if the thermal resistance is the same for both, the power dissipated in the driver stage is much lower than in the output stage. This is also revealed by Figure 3.45 where the simulated DC current of the driver stage is lower than the measured one.

It is then possible to restore the desired performance of this power amplifier by a proper design of the output network (eventually with the use of a load pull instrument to synthesize it) and an improved thermal circuit able to reduce the temperature variation into the die.

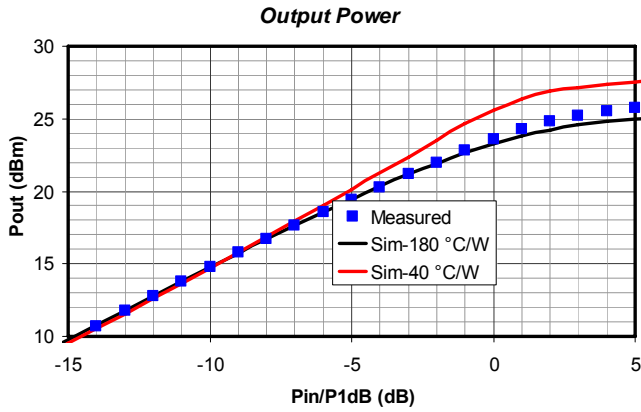


Figure 3.42 Measured and simulated output power

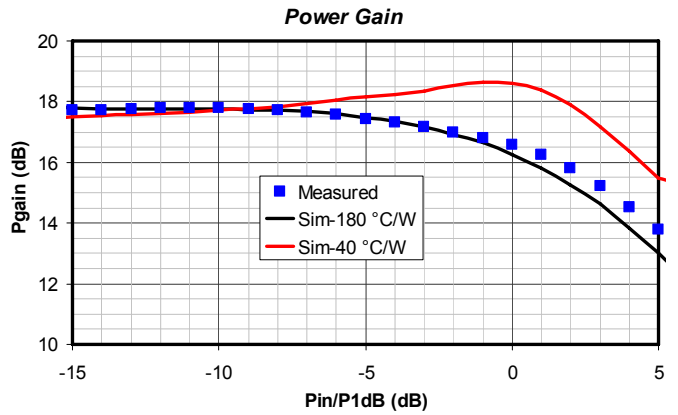


Figure 3.43 Measured and simulated power gain

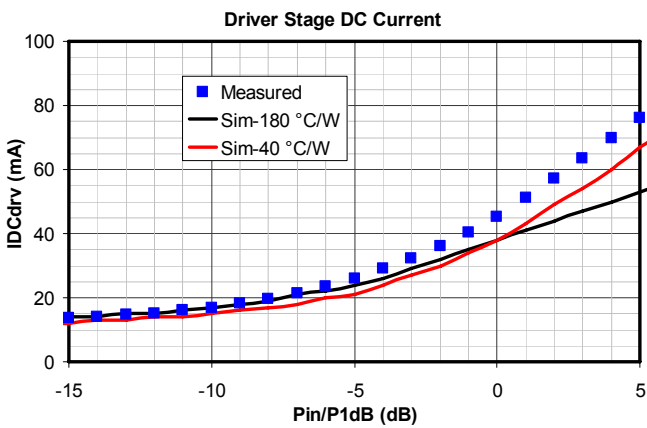


Figure 3.45 Measured and simulated driver stage DC current

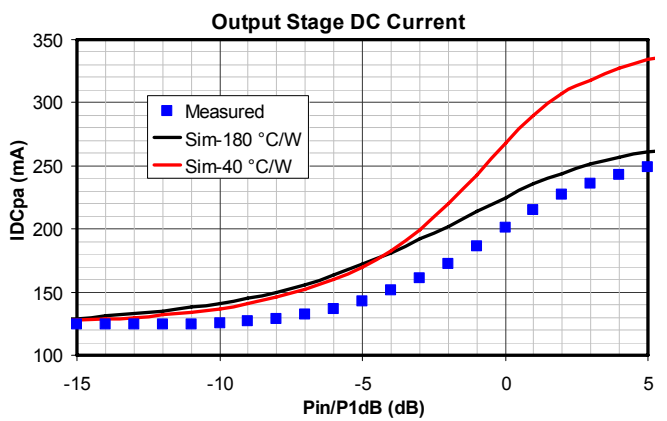


Figure 3.44 Measured and simulated output stage DC current

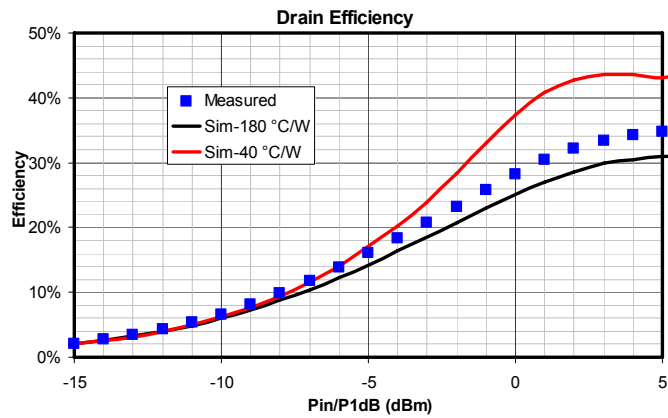


Figure 3.46 Measured and simulated drain efficiency

3.4 Conclusion

In this chapter the design and testing of a common base class AB power amplifier has been presented. This topology is able to increase the efficiency of a standard cascode topology while maintaining its compliance on avalanche breakdown current. This allows to bias the power amplifier above BV_{CEO} thus allowing to increase the load impedance. The inter-stage network between the driver stage and the output stage allows to reduce the size of the former thus reducing its dissipated power. A off chip impedance transformation network has been implemented. The measured output power and efficiency are lower than the expected because of the action of the testing board in terms of output load and thermal effects. That notwithstanding the working principle of this structure has been demonstrated, since the measurement are influenced by the off chip elements.

References

- [12] D. T. S. Cheung, and J. R. Long, “*A 21-26 GHz SiGe Bipolar Power Amplifier MMIC*”, IEEE Journal of Solid-State Circuits, vol. 40, No. 12, pp 2583-2597, Dec 2005
- [13] H. Veenstra, G. A. M. Hurkx, D. van Goor, H. Brekelmans, and J. R. Long, “*Analyses and Design of Bias Circuits Tolerating Output Voltages Above BVCEO*”, IEEE Journal of Solid-State Circuits, vol. 40, NO. 10, pp2008-2018, Oct 2005
- [14] A. Scuderi, L. La Paglia, A. Scuderi, F Carrara, and G. Palmisano, “*A VSWR-Protected Silicon Bipolar RF Power Amplifier with Soft-Slope Power Control*”, IEEE Journal of Solid-State Circuits, vol. 40, NO. 3, pp 611-621, Mar 2005
- [15] M. Rickelt, H.-M. Rein, and E. Rose, “*Influence of impact-ionization induced instabilities on the maximum usable output voltage of Si-bipolar transistors*” IEEE Trans. Electron Devices, vol. 48, no. 4, pp. 774–783, Apr. 2001
- [16] A. Inoue, S. Nakatsuka, R. Hattori, and Y. Matsuda, “*The maximum operating region in SiGe HBTs for RF power amplifiers*” in IEEE MTT-S Int. Microwave Symp. Dig., vol. 2, pp. 1023–1026, Jun. 2002
- [17] Avanzo, F.; De Paola, F.M.; Manstretta, D.; “*A common-base linear rf power amplifier for 3G cellular applications*” Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE 21-24 Sept. 2008 Page(s):579 – 582.

Chapter 4

The Design of a CMOS Doherty Power Amplifier

The Doherty Efficiency enhancement technique seems to be the most straightforward way to enhance the efficiency of a linear power amplifier. The goal is to maintain the efficiency of a linear power amplifier close to the maximum value on a wider power range. In other techniques (like EER and Chireix) a non linear but efficient power amplifier is used and the goal is to conveniently restore the amplitude modulation at the output. Thus it seems natural to use a linear PA and enhance its efficiency instead to employ a non-linear PA surrounded by other (non efficient) elements that restore the amplitude variations.

In this chapter the design of a Doherty PA is discussed, posing the attention to the Auxiliary PA, which has the major role in the structure performance. A solution which allows to employ a modulator which conveniently varies the phase shift at the signal apply to the main and the peaking amplifier will be considered, allowing to integrate the overall transmitter.

4.1 The Ideal Doherty Structure

The Doherty PA (which basic idea based on a dynamic impedance variation has been presented in Pararaph 2.3.1 and reported again in Figure 4.1) consists of a Main Amplifier and Auxiliary Amplifier which operates a dynamic variation of the resistance seen at the main amplifier's output.

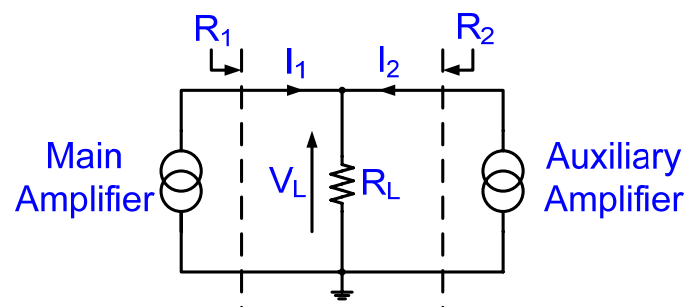


Figure 4.1 Dynamic impedance variation

In a Doherty structure, two different operative regions can be identified: when the Aux is off and when it is on. The point where the auxiliary amplifier is turned on ($V_{IN-AuxOn}$) represents the input voltage (power) where the efficiency is wanted to be maximized. Generally it is placed at a 6dB back off from the P_{1dB} . In the upper range the auxiliary amplifier is turned on, and the main amplifier is working at its maximum voltage efficiency, i.e. the output voltage is kept at the maximum swing. In this range the auxiliary amplifier contributes to the output power in order to compensated the reduced power supplied by the main amplifier. The ideal RF voltage and currents at the main and auxiliary amplifier output are reported in the next figure.

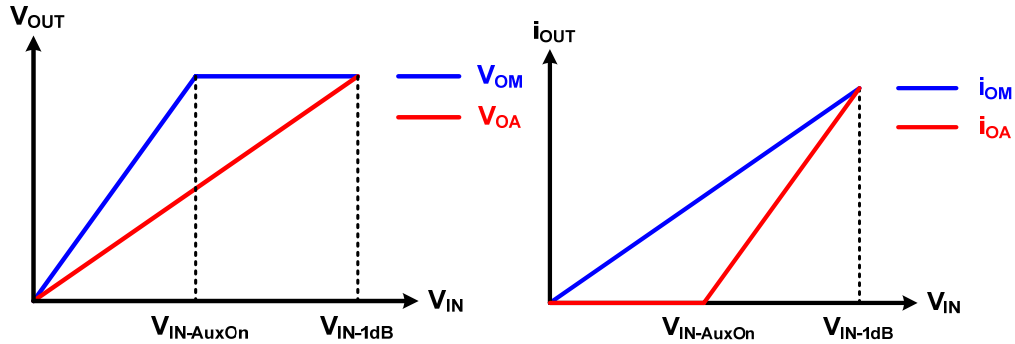


Figure 4.2 Output RF voltages and currents of an ideal Doherty Amplifier

It is possible to see that the auxiliary amplifier shows a strongly nonlinear characteristic. Hence, in order to preserve a linear I/O characteristic this current should not flow into the load. The block diagram of the Doherty amplifier which implements the desired voltage and current behavior giving the desired efficiency is reported in Figure 4.3 [18].

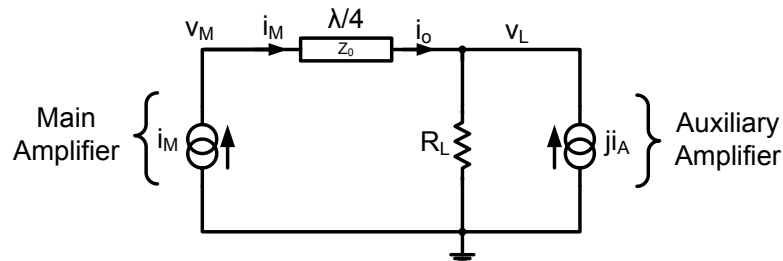


Figure 4.3 Block diagram of an ideal Doherty PA

The quarter-wave line at the Main PA works as an impedance transformer acting as an impedance inverter, which causes the resistive impedance seen by the main device to go down as the auxiliary device current i_A increases. Moreover this quarter-wave line prevents the auxiliary output current from flowing into the load, giving a low impedance seen at the auxiliary output (which is in parallel to the load). In fact, the impedance seen at the input of a quarter-wave line is equal to Z_0^2/R_L where Z_0 is the characteristic impedance of the line and R_L is (in this case) given by the output resistance of the main amplifier: since it is very high, the impedance seen at the other side is very low.

In order to calculate the expressions of the voltage seen at the main amplifier output (V_{OM}), let's first recall the matrix which links the input and output voltages and currents in a quarter-wave line (referring to the quantities in Figure 4.3):

$$\begin{bmatrix} v_M \\ i_M \end{bmatrix} = \begin{bmatrix} 0 & jZ_0 \\ \frac{j}{Z_0} & 0 \end{bmatrix} \cdot \begin{bmatrix} v_L \\ i_0 \end{bmatrix} \quad 4.1$$

The transmission matrix determines the following relations:

$$v_M = jZ_0 i_0 \quad \Rightarrow \quad i_0 = -j \frac{v_M}{Z_0} \quad 4.2$$

$$i_M = j \frac{v_L}{Z_0} \quad \Rightarrow \quad v_L = -jZ_0 i_M \quad 4.3$$

Applying the KCL at the output node we get:

$$\frac{v_L}{R_L} = i_0 - j i_A \quad 4.4$$

Replacing the 4.2 and 4.3 in the previous equation we obtain:

$$\begin{aligned} -j \frac{Z_0}{R_L} i_M = -j \frac{v_M}{Z_0} - j i_A & \Rightarrow \quad \frac{v_M}{Z_0} = \frac{Z_0}{R_L} i_M - i_A \quad \Rightarrow \\ \Rightarrow \quad v_M = Z_0 \left[\left(\frac{Z_0}{R_L} \right) i_M - i_A \right] & \quad 4.5 \end{aligned}$$

This equation gives what we are looking for: the possibility to keep constant the voltage at the main amplifier output due to the action of the auxiliary amplifier. This allows the main amplifier to work with a maximum voltage efficiency. Moreover the action of the auxiliary amplifier doesn't affect the voltage on the load, as stated by 4.3.

The auxiliary amplifier turns on only when the input voltage exceeds a threshold value. The slope of the output current of the auxiliary needs has to be high enough to equal the main amplifier current at the maximum power delivered. Since the auxiliary is generally turned on when the output power is 6dB below the P_{1dB} , the input voltage at this point will be halved compared to the maximum input voltage.

Since the main amplifier output voltage depends on R_L and Z_0 it is necessary to calculate their value which allows to get the desired performance. In order to do this, let's consider the maximum output current of the main and the auxiliary amplifier:

$$(i_M)_{\max} = i_{M,\max} \quad (i_A)_{\max} = i_{A,\max}$$

The input voltage when the current starts to flow in the auxiliary amplifier is $v_{IN-AuxOn}$ (also called breakpoint). Its value, normalized to the maximum input voltage, is equal to 0.5. In order to allow the main amplifier output voltage to reach its prescribed maximum level, the R and Z_0 values must be conveniently chosen. If we suppose that the auxiliary amplifier is off at the breakpoint, we want the main amplifier to show at this point an RF voltage with an amplitude V_{DC} (equal to its supply voltage); its current has to be $I_{MAX}/4$ where I_{MAX} is the maximum RF current which flows into the load at P_{1dB} . Under these conditions the 4.5 becomes:

$$4.6 \quad @ \text{ breakpoint :} \quad v_M = V_{DC} = 0.5 \frac{Z_0^2}{R_L} i_{M,max}$$

When the maximum input voltage is applied, the auxiliary output current will equal the main RF current (which is $I_{MAX}/2$ since it contributes for half of the output power), and the main amplifier RF voltage must be V_{DC} in order to have maximum voltage efficiency. Thus we obtain:

$$4.7 \quad @ v_{IN,max} : \quad v_M = V_{DC} = Z_0 \left[\left(\frac{Z_0}{R_L} \right) i_{M,max} - i_{M,max} \right]$$

Equating 4.6 and 4.7 we obtain:

$$4.8 \quad R_L = \frac{V_{DC}}{2i_{M,max}} \quad Z_0 = \frac{V_{DC}}{i_{M,max}}$$

Thus the characteristic impedance of the quarter-wave line must be twice the load impedance in order to let the main amplifier to reach the maximum RF voltage at the breakpoint. When the auxiliary amplifier is off, the impedance seen at the main amplifier output is four times the load impedance. This result can be intuitively explained considering that the main amplifier must reach its maximum swing 6dB below the P_{1dB} that means a quarter of the maximum output power.

With this topology it is possible to have a big benefit in the efficiency behavior compared to a common linear power amplifier. The performance comparison in terms of efficiency are shown in Figure 4.4. The main advantages of a Doherty structure are visible in the region before the breakpoint. Here the efficiency is almost doubled compared to the equivalent Class B amplifier. This because the main amplifier is designed to reach its maximum efficiency when the power is 6dB lower (4 times smaller) than the single stage case. Since the efficiency of a Class B amplifier decreases with a $\sqrt{2}$ slope, the efficiency gain of a Doherty structure in the auxiliary-off range is exactly a factor of two. Another remarkable advantage of this structure is that the action of the auxiliary non linear amplifier is invisible at the load since the dependency of the output power on the input drive signal remains defined by the main device characteristic, which can be much more linear. This is not actually true in a real case where the main device has a finite output impedance. This issue is discussed in the next paragraph.

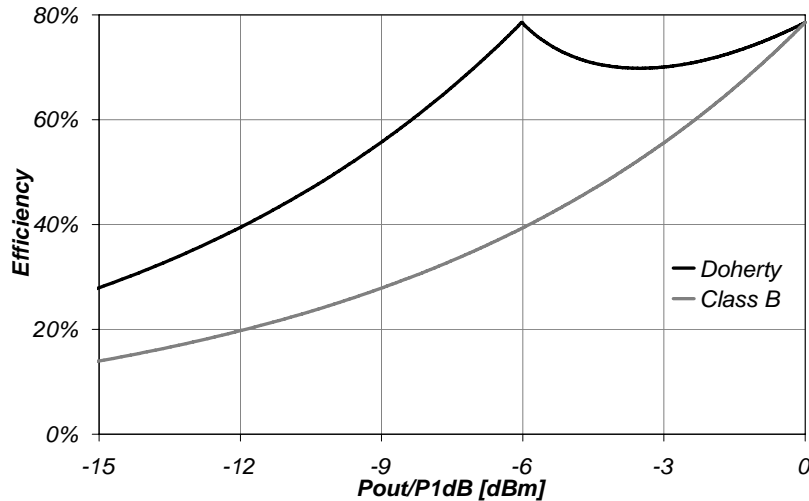


Figure 4.4 Efficiency behavior of an ideal Doherty PA compared to a Class B Amplifier.

4.2 Second order effects

The main second order effects, limiting the performance of an actual Doherty structure in terms of linearity and efficiency, are related to the finite output resistance of the amplifiers and the losses due to the quarter-wave line. Another source of performance reduction is related to an incorrect phase shift between the currents of the main and auxiliary amplifiers.

4.2.1 Finite output resistance

In the previous paragraph the equation for an ideal Doherty PA has been derived, showing the insensitivity of the load voltage (v_L) from the nonlinear behavior of the auxiliary amplifier. In that case the amplifiers considered were ideal current sources with infinite output impedance. In order to predict a more realistic behavior, let's redraw the scheme of Figure 4.3 with the output resistances of the main and the auxiliary amplifier (R_{OM} and R_{OA}) as in Figure 4.5.

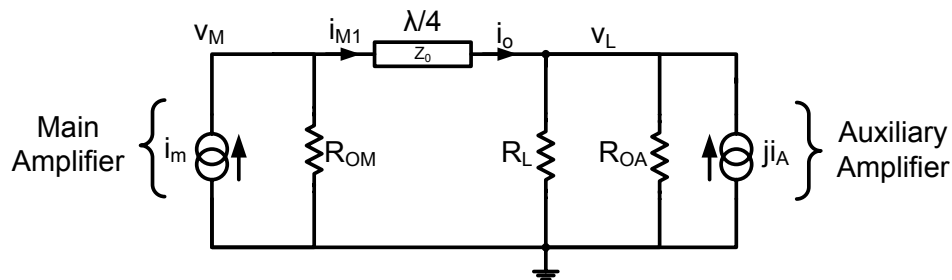


Figure 4.5 Ideal Doherty structure with finite output resistance

The effect of finite output impedances is, from one side, to reduce the load impedance (since R_{OA} is in parallel with R_L) and from the other side to increase the impedance seen by the auxiliary amplifier (which is not anymore null). This allows part of the auxiliary current to flow into the load impedance. Hence, the output voltage will show some nonlinearity.

In fact the load voltage is now described by:

$$4.9 \quad v_L = -jZ_0 i_{M1} \Rightarrow v_L = -jZ_0 \left(i_M - \frac{v_M}{R_{OM}} \right)$$

where

$$4.10 \quad v_M = jZ_0 i_O \Rightarrow v_M = Z_0 \left(j \frac{v_L}{R_{Leq}} - i_A \right)$$

R_{Leq} is the parallel connection between R_L and R_{OA} . Due to the action of R_{OM} the load voltage is now dependent by i_A via v_M . The full expression of the load voltage given by rearranging 4.9 and 4.10 is:

$$4.11 \quad V_L = -jZ_0' \left(i_M + \frac{Z_0}{R_{OM}} i_A \right)$$

Where

$$4.12 \quad Z_0' = \frac{Z_0}{1 + \frac{Z_0^2}{R_{OM} R_{Leq}}}$$

which becomes the 4.3 when R_{OM} becomes infinite. Figure 4.6 shows the behavior of the normalized load voltage for the normalized main current for different amounts of the output resistance. It is possible to see that a nonlinear shape appears when the output resistance becomes less than 25 times compared to the load impedance. This effect plays an important role in the design of the main amplifier, whose topology choice will be influenced by this effect.

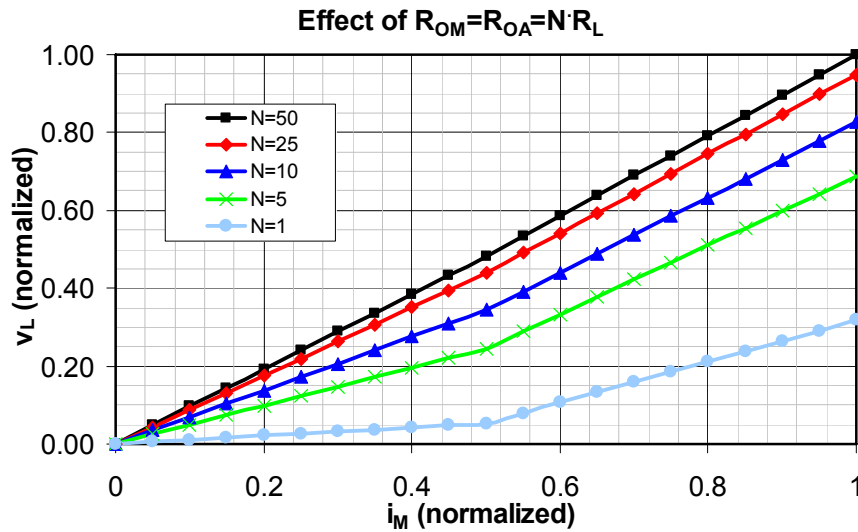


Figure 4.6 Effect of the main amplifier output impedance on the load voltage

4.2.2 Phase mismatch

As discussed in Chapter 3, the impedance reduction due to the effect of the auxiliary amplifier takes place if a proper phase shift between the main and auxiliary amplifier currents is present. If this phase shift is not correct, a wrong impedance transformation takes place, and it causes nonlinearity and efficiency degradation.

The 4.11 has been obtained assuming a -90° phase shift in the auxiliary amplifier output current. If we now assume to have a generic phase shift in i_A the 4.11 becomes:

$$V_L = -Z_0 \left(j i_M - \frac{Z_0}{R_{OM}} \bar{i}_A \right) \quad 4.13$$

Where \bar{i}_A represents a current with a generic phase shift. In Figure 4.7 and Figure 4.8 the behavior of the load voltage v_L magnitude and phase for several phase shift is reported supposing to have an output resistance 25 times larger than the load. It is possible to see that a phase shift different from the ideal -90° acts in two ways: it can increase or reduce the phase variation at the output voltage.

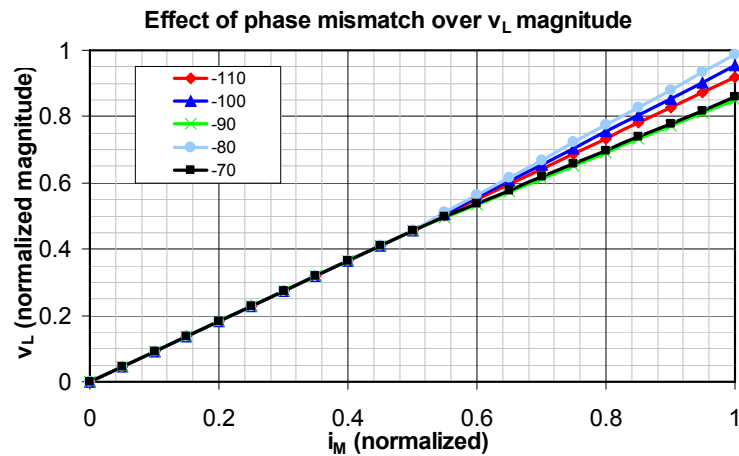


Figure 4.7 Effect of phase mismatch to the magnitude on the load voltage

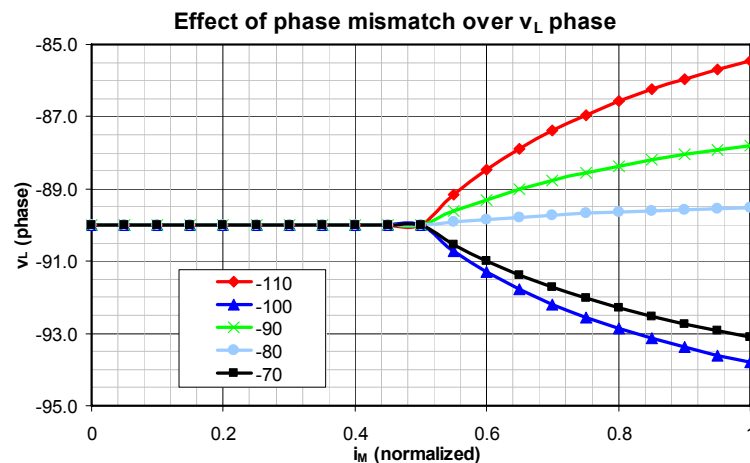


Figure 4.8 Effect of phase mismatch to the phase on the load voltage

Looking at the curve relative to the -80° phase shift it is possible to note a reduction of the phase variation with the amplitude of the main amplifier current.

4.2.3 Quarter-wave line

So far no comments has been made about the realization of the quarter-wave line. In fact its realization depends on the application where the Doherty amplifier is going to be used. The UMTS standard works in uplink at around 1.95GHz: if an integrated solution of the Doherty structure is wanted, on-chip microstrip quarter-wave line is not feasible, since it would have a length of several millimeters. Thus, for RF applications, it is possible to implement a quarter-wave line with lumped parameters circuits. Two different implementations are shown in Figure 4.9, respectively high pass (a) and low pass (b).

The two pi-network synthesize the quarter-wave line with the following transmission matrixes (respectively for figure a and b):

$$4.14 \quad \begin{bmatrix} v_1 \\ i_1 \end{bmatrix} = \begin{bmatrix} 0 & jZ_0 \\ \frac{j}{Z_0} & 0 \end{bmatrix} \cdot \begin{bmatrix} v_2 \\ i_2 \end{bmatrix} \quad \begin{bmatrix} v_1 \\ i_1 \end{bmatrix} = \begin{bmatrix} 0 & -jZ_0 \\ -\frac{j}{Z_0} & 0 \end{bmatrix} \cdot \begin{bmatrix} v_2 \\ i_2 \end{bmatrix}$$

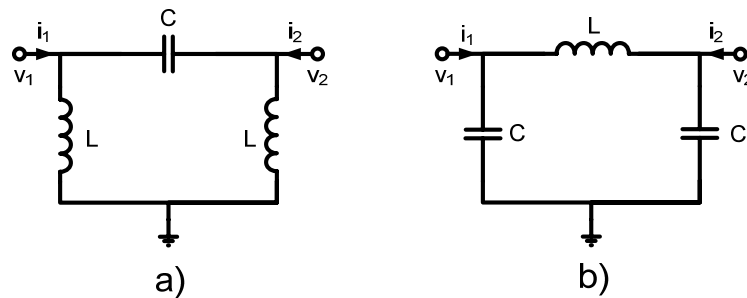


Figure 4.9 High pass a) and low pass b) concentrate parameters impedance inverters

where

$$4.15 \quad Z_0 = \sqrt{\frac{L}{C}} \quad @ \omega_0 = \frac{1}{\sqrt{LC}}$$

The main difference between the two realizations regards the phase shift which is 90° in the first case and -90° in the second. Thus a proper phase of the auxiliary amplifier must be set depending on the network used. The choice between the two realizations mainly depends on layout considerations that will be discussed later. These networks introduce losses which are related to the quality factor of the inductors (which generally have lower Q if compared to the integrated capacitors). Thus high Q inductors must be used in order to limit the efficiency reduction.

4.3 Doherty Power Amplifier Design

The design of a complete Doherty structure is now presented, with emphasis on the auxiliary amplifier implementation, a critical element which limit the performance of the structure. The design starts from the main amplifier, which is optimized in terms of linearity and efficiency. The design flow of this amplifier was explained in Paragraph 3.3.1. Once the load resistance is chosen in order to supply the desired output power, the quarter-wave line and the auxiliary amplifier constraint can be obtained. The aim is to obtain a Doherty power amplifier able to deliver a 30dBm signal power at the 1dB compression point to a 50Ω load antenna with sufficient linearity and optimized efficiency. The technology considered is a 65nm CMOS.

4.3.1 Main Amplifier Design and Impedance Transformer.

The main amplifier design starts with the design of a single-stage amplifier able to deliver the maximum power of the overall Doherty structure. This will be either a starting point for the main amplifier design and also a term of comparison for the efficiency performance between the Doherty structure and a classical linear amplifier. Once this amplifier is optimized the main amplifier is obtained by conveniently scaling its size, since the load impedance seen at its output is made larger due to the action of the quarter-wave line.

First, the main amplifier topology must be chosen since it plays an important role in the linearity performance of a Doherty power amplifier. In fact, as already introduced in Paragraph 4.2.1, the output resistance of the main amplifier should be at least 25 times larger than the load impedance, in order to limit the nonlinear behavior. Among the topologies shown in paragraph 3.2, the cascode solution shows the higher output resistance compared to the simple common emitter and common base amplifier, although it has a lower efficiency compared to the latter. It must be noted that in this case the need for low gate impedance to avoid avalanche breakdown is not present because of the CMOS technology used.

In the next paragraphs a single end solution will be considered, while the pseudo differential complete solution will be shown in paragraph 4.3.5. This single end solution will be designed to obtain a 27dBm maximum linear power with a 2.5V power supply. The pseudo-differential solution will gain a factor of two in the output power, thus achieving the desired output level.

The amplifier designed following the flow described in paragraph 3.3.1 is reported in Figure 4.10. The load resistance R_L is obtained by an impedance transformation from the 50Ω antenna, and its value is 4.5Ω . The L_C inductor resonates out the output capacitance (C_{OUT}) of M_2 , which is 6.4pF and thus needs 1nH to be resonated at 1.95GHz. The performance of this amplifier are reported in Figure 4.11- Figure 4.15. The aim of L_2 and C_2 is to provide a low path at the second harmonics of the drain current.

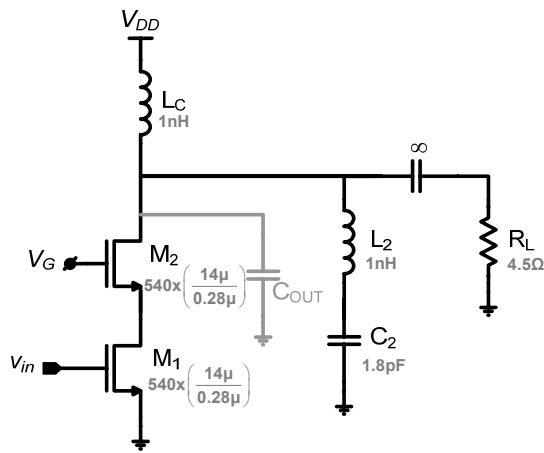


Figure 4.10 Class AB power amplifier

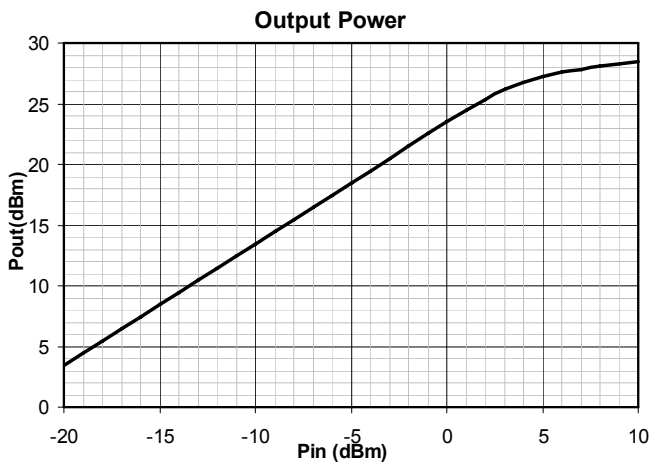


Figure 4.12 Output Power of the Class AB amplifier

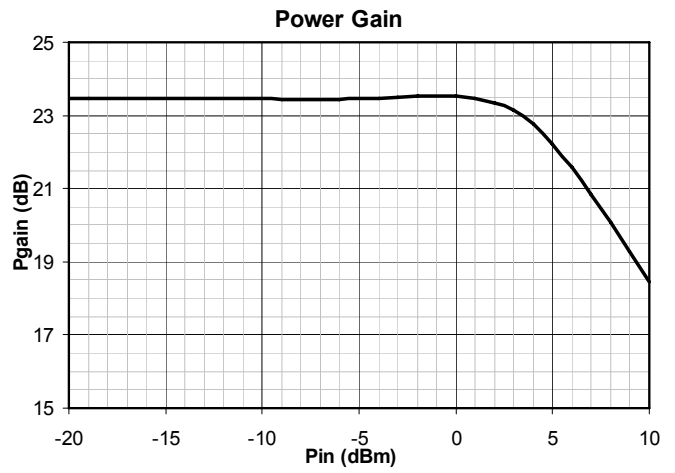


Figure 4.11 Power Gain of the Class AB amplifier

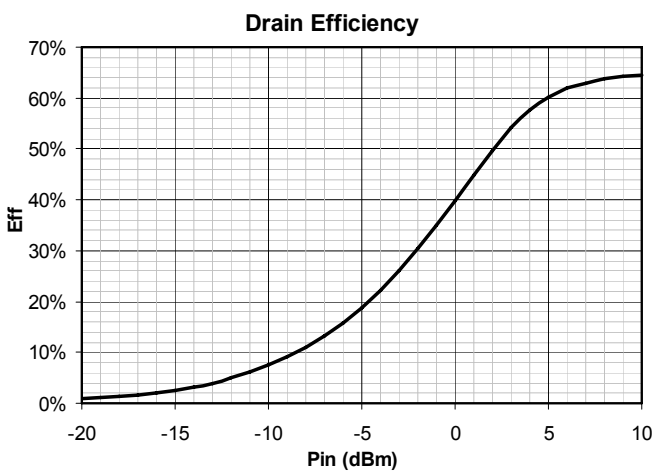


Figure 4.14 Drain Efficiency of the Class AB amplifier

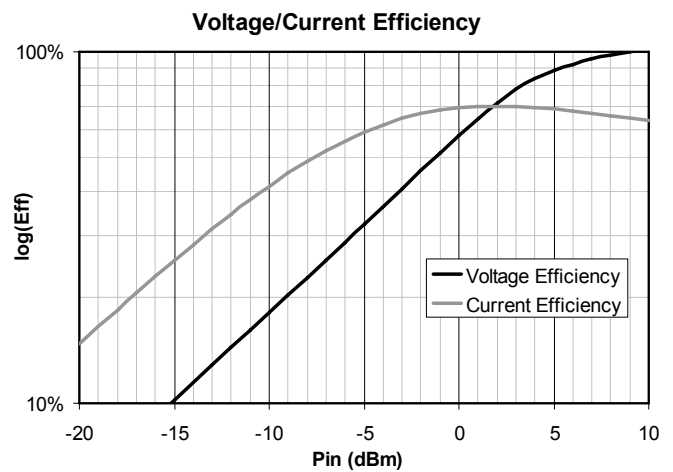


Figure 4.13 Voltage and current efficiency of the Class AB amplifier

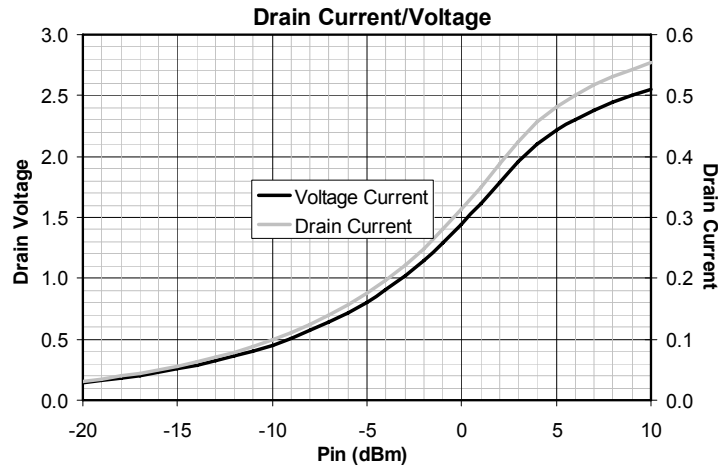


Figure 4.15 Drain voltage and current of the Class AB amplifier

The amplifier has an output 1dB compression point of around 27dBm (Figure 4.12) with an input 1dB compression point of 5dB (Figure 4.11). The maximum linear drain efficiency is 60% (Figure 4.14). For what concerns the voltage and current efficiencies, the former reaches the 90% at the compression point (because of the transistor's saturation voltage), while the current efficiency is constant until 7dB of back off. This is an useful behavior in order to scale this amplifier to implement the main amplifier. Since the scaling of the transistor (with an equivalent increase on the load impedance) doesn't change the performance in terms of efficiency (drain, voltage and current), this behavior of the current efficiency suggests that a proper increase in the voltage efficiency will increase the overall efficiency. If the drain efficiency at the back off would have been lower than at the P_{1dB} , an increase in the voltage efficiency would not allow to the drain efficiency to reach its maximum value.

Now that the starting point is set, the quarter-wave line must be implemented and the transistor scaled. The choice of the quarter-wave line topology is mainly based to layout and area considerations, since (apart from the phase shift introduced) the two networks in Figure 4.9 have equivalent performance. Figure 4.16 shows a possible implementation of the Doherty structure with the main and the auxiliary amplifier. It is possible to see that the implementation of the impedance inverter with the solution of Figure 4.9a helps to reduce the area. In fact, the inductors of the quarter-wave line are in parallel with the inductors L_1 and L_2 which are necessary in order to bias the transistors and resonate the output capacitance (here the transistors are not cascoded for simplicity, since this is a general solution). In this way the quarter-wave line can simply implemented in the way depicted in Figure 4.17. A proper sizing of L_A , L_B (which are obtained by the parallel connection of the drain inductor and that of the quarter wave line) and C allows to implement the impedance inverter just with a "smart" connection of the two amplifiers by a simple capacitor. The other solution would have required another inductor for the connection between the two stages, thus increasing the occupied area.

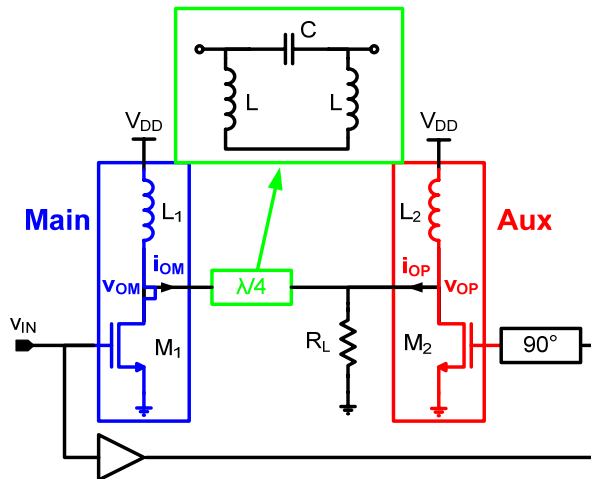


Figure 4.16 Single ended Doherty amplifier implementation

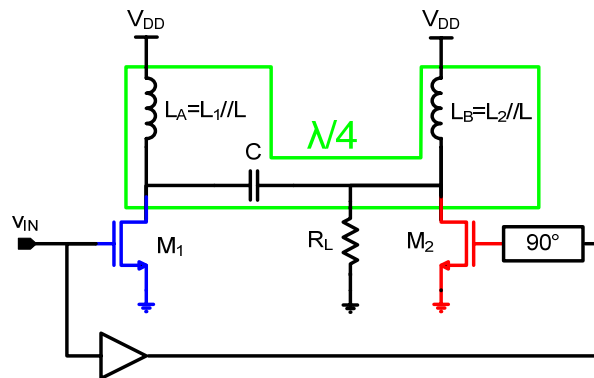


Figure 4.17 Realization of the impedance inverter structure

Since the load resistance is 4.5Ω and the characteristic impedance of the quarter wave line must be 2 times higher (equation 4.8), the L and C value should be (in order to resonate at 1.95GHz):

$$4.16 \quad C = \frac{1}{2\pi f_0 Z_0} = 9.1\text{pF}$$

$$4.17 \quad L = Z_0^2 C = 734\text{pH}$$

Since the load impedance seen at the main amplifier output is equal to $4R_L$ due to the effect of the impedance inverter, if the maximum efficiency is wanted at 6dB below the $P_{1\text{dB}}$, the transistor's size should be 4 times smaller compared to the solution of Figure 4.10. This is a different approach compared to the classic implementation, where the "starting" amplifier is split in two amplifiers (main and auxiliary) having the same size (equal to the half of the single stage amplifier). With the classic approach the auxiliary amplifiers delivers the same current at the $P_{1\text{dB}}$ as the main amplifier, but the efficiency is not maximized at the breakpoint. With the approach chosen here the auxiliary amplifier will be larger than the main (in order to compensate the lower current delivered by this stage) but the efficiency will be maximized at the turn-on. The quantitative comparison will be shown in the next paragraph.

4.3.2 Ideal Auxiliary Amplifier

The ideal behavior of the auxiliary amplifier drain current (i_{oA}) is shown in Figure 4.18. The auxiliary turns on at $V_{IN-AuxOn}$ with an abrupt change of the supplied current: in this way it is possible to obtain the maximum achievable efficiency performance for the Doherty structure at the breakpoint. This will be the term of comparison among different realizations of the auxiliary amplifier in order to determine the entity of efficiency reduction due to a more realistic implementation.

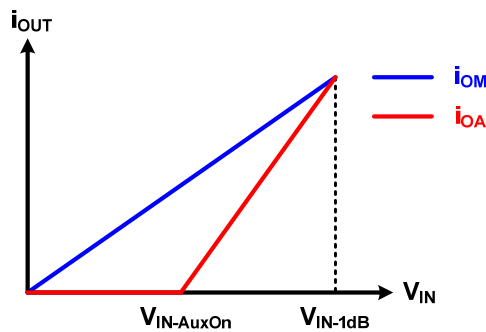


Figure 4.18 Ideal current behavior of the main and auxiliary amplifiers

This implementation of the auxiliary amplifier allows also to compare the performance for different sizes of the main amplifier (as discussed in the previous paragraph). The next figures show the performance comparison between the classical Doherty implementation (where the main amplifier has half of the size of the single stage cascode Figure 4.19) and the proposed solution (where its size is a quarter of the single stage amplifier Figure 4.20) and a single stage class AB implementation.

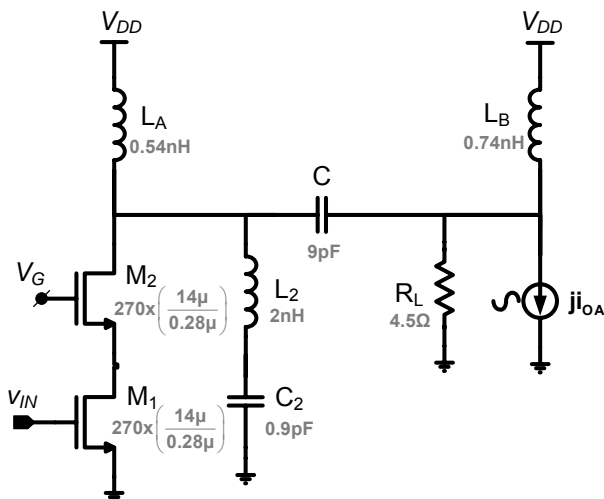


Figure 4.20 Doherty Amplifier with classical amplifier sizing

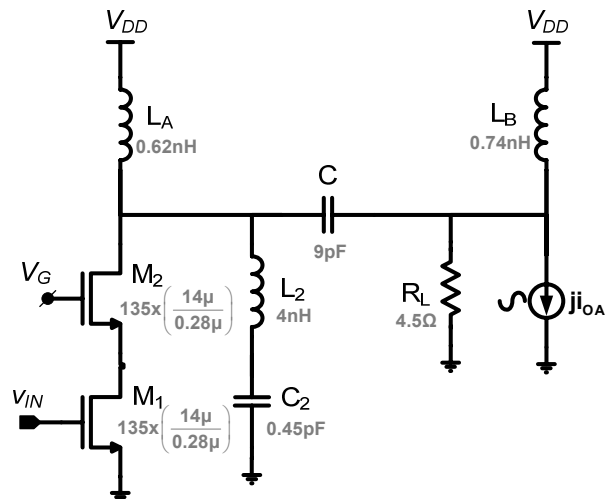


Figure 4.19 Doherty Amplifier with optimized amplifier sizing

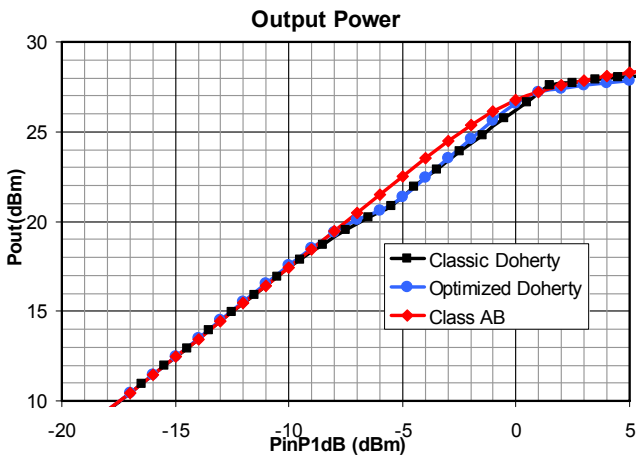


Figure 4.21 Simulated output power comparison

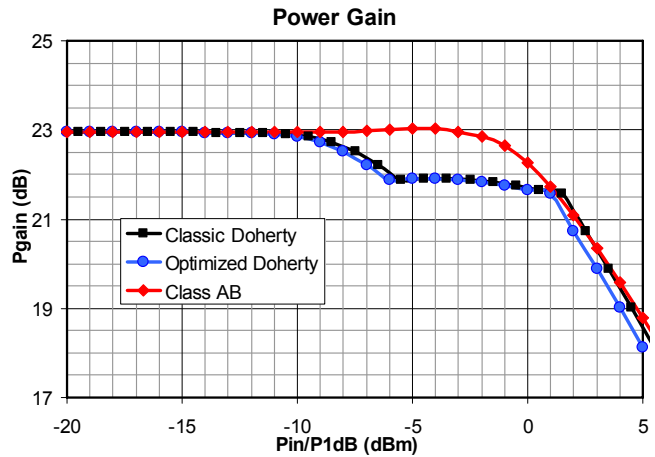


Figure 4.22 Simulated power gain comparison

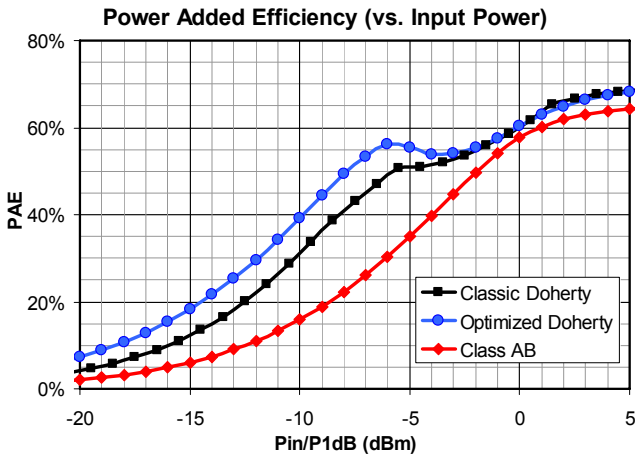


Figure 4.24 Simulated PAE (vs. input power) comparison

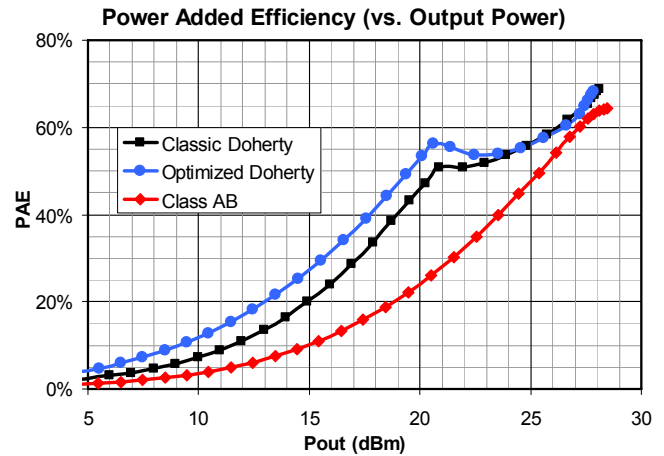


Figure 4.23 Simulated PAE (vs. output power) comparison

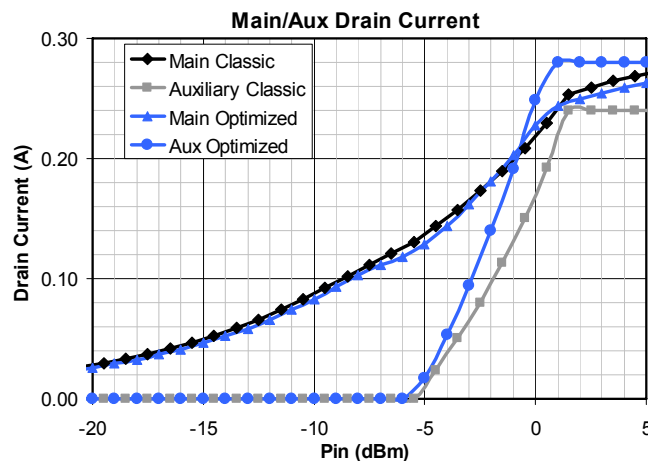


Figure 4.25 Simulated output current comparison

The optimized solution allows to gain some efficiency (around 20% at the turn on and 30% at the lower power levels) at the power levels before the breakpoint (Figure 4.24 and Figure 4.23). Moreover the main amplifier has half of the size compared to the classical solution. The reduced size is compensated by a higher

current supplied the auxiliary amplifier (Figure 4.25) which will need a bigger device area. Thus the difference between the classical and the optimized solution relates principally to the improved efficiency than a reduced area.

This solution with an ideal auxiliary amplifier gives the maximum performance in terms of efficiency (especially at the breakpoint), but determines a nonlinear behavior in the power gain (Figure 4.22). This happens because the auxiliary turns on at the compression point of the main amplifier, where its efficiency is maximized but the power gain is reduced. Thus, in order to avoid anticipated gain compression, the auxiliary amplifier should be turned on at a lower power level, where the main amplifier efficiency is not maximized. This reveals a trade-off between linearity and maximum efficiency at the breakpoint.

The advantage of the optimized main amplifier in terms of efficiency compared to the classical solution is more evident (Figure 4.27) when the auxiliary amplifier turn-on is chosen to minimize distortion of the power gain (Figure 4.26). Although the efficiency at the breakpoint is lower compared to the previous case, the benefit of the power gain behavior lets the Doherty structure to be as linear as the single stage solution, while maintaining higher efficiency over a large range of output power. The next paragraph will show a practical realization of the auxiliary amplifier biased below the threshold (class C).

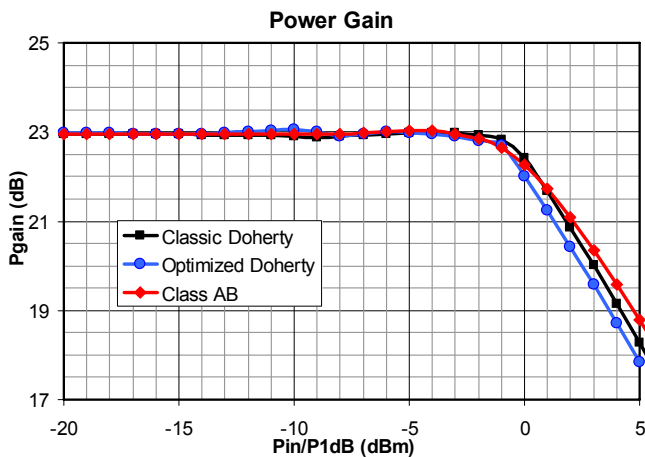


Figure 4.26 Power Gain with anticipated breakpoint

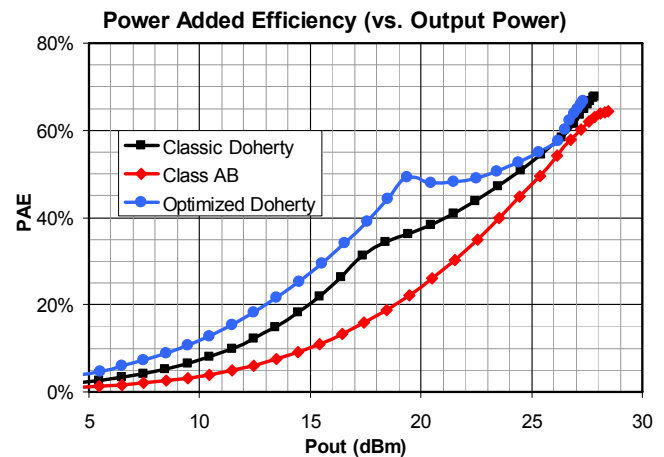


Figure 4.27 PAE with anticipated breakpoint

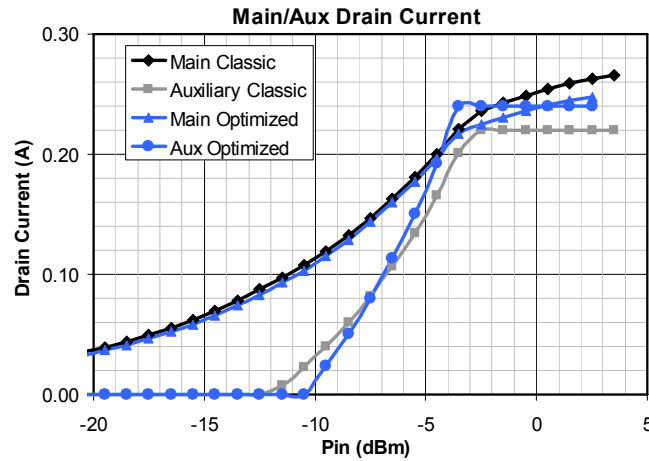


Figure 4.28 Drain Current with anticipated breakpoint

4.3.3 Class C Auxiliary Amplifier

The ideal behavior of the auxiliary drain current is difficult to replicate with a single stage amplifier, because of the abrupt change of the supplied current. The classical way to realize a similar behavior is the use of a Class C amplifier.

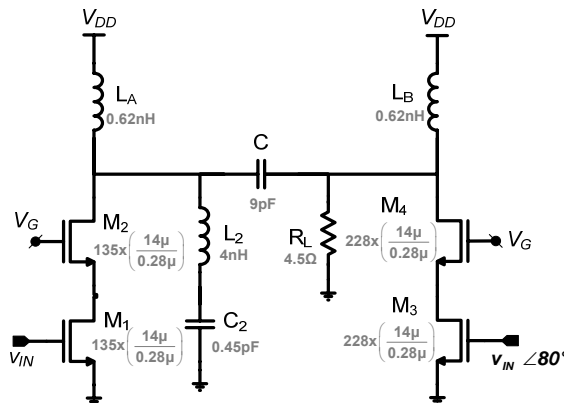


Figure 4.29 Doherty Amplifier with Class C auxiliary

In Class C, since the transistor is biased below its threshold, conduction takes place only when the input signal is sufficiently high, thus the drain current shows an impulsive behavior. The shape of the RF drain current over the input signal recalls the nonlinear current behavior considered for the auxiliary amplifier in the previous paragraph but with a less sharp turn on.

In order to correctly replicate a current behavior for the auxiliary amplifier, a proper device size and bias voltage shall be chosen. The use of a class C amplifier will reduce the efficiency at the auxiliary turn on compared to the ideal solution, since the drain current will not show an “instantaneous” turn non behavior. The performance of a Doherty structure which uses a class C auxiliary amplifier is shown in the next figures, and its performance are compared to the ideal solution already considered.

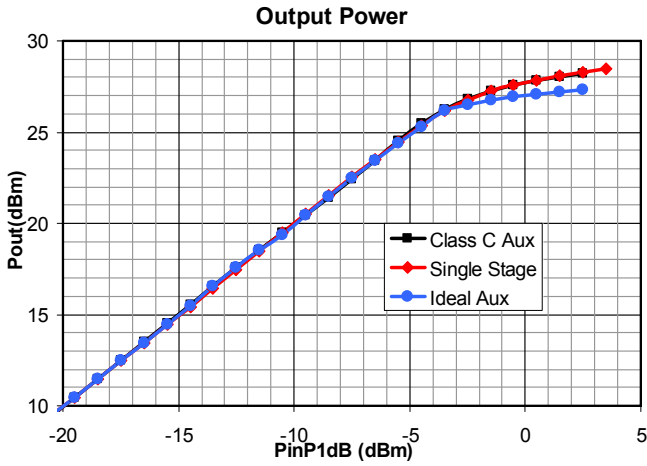


Figure 4.30 Simulated output power with Class C Auxiliary amplifier

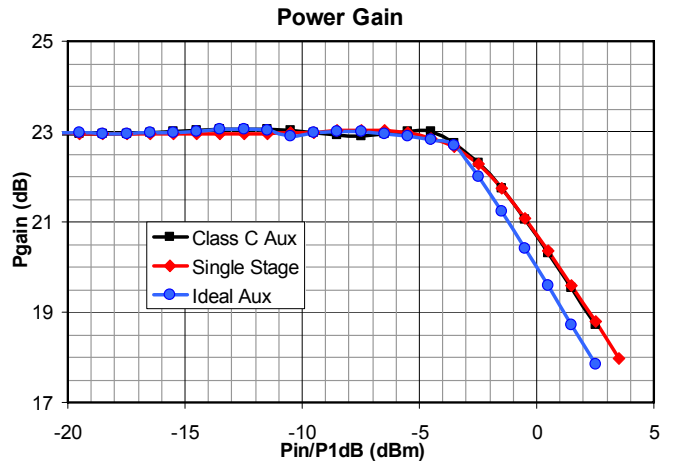


Figure 4.31 Simulated power gain with Class C Auxiliary amplifier

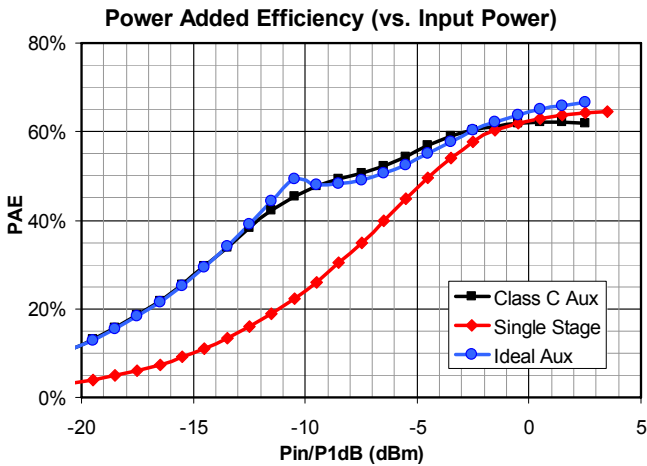


Figure 4.33 Simulated PAE (vs. Input power) with Class C Auxiliary amplifier

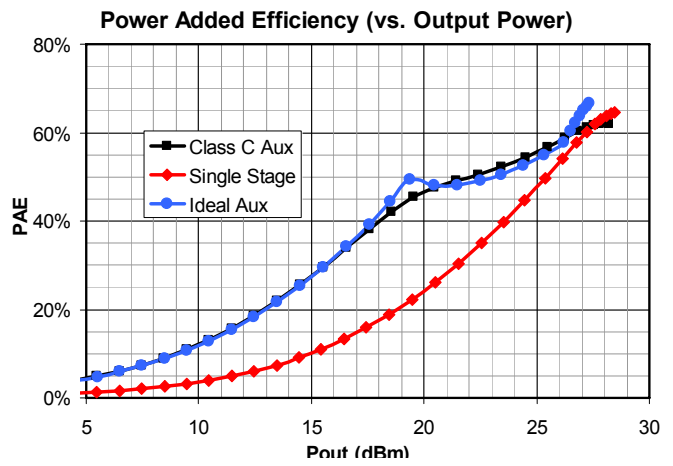


Figure 4.32 Simulated PAE (vs. output power) with Class C Auxiliary amplifier

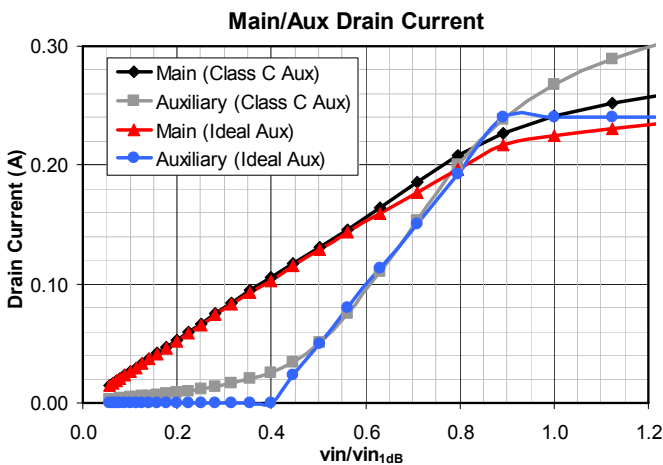


Figure 4.35 Simulated Main/Aux drain current with Class C Auxiliary amplifier

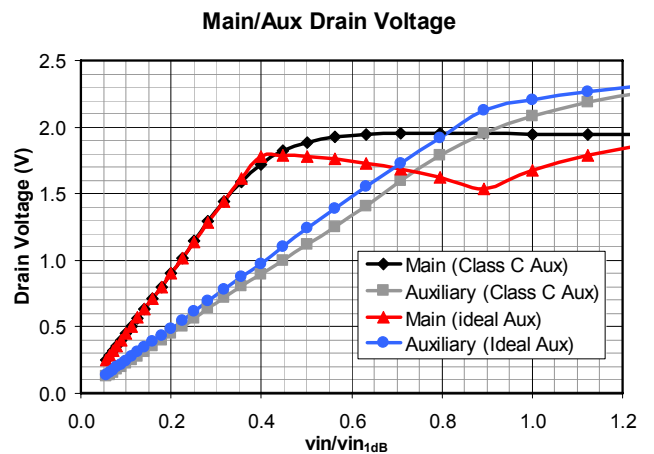


Figure 4.34 Simulated Main/Aux drain voltage with Class C Auxiliary amplifier

It is possible to see from the previous graphs that a Class C auxiliary amplifier doesn't introduce a dramatic reduction in the efficiency, especially at the "turn on" (which is now more difficult to identify looking at the efficiency). The output power and power gain (Figure 4.31 and Figure 4.30) has the same behavior as the single stage class AB solution, while the efficiency (Figure 4.32 and Figure 4.33) has a remarkable improvement. Thus a class C stage is a good choice to implement the auxiliary amplifier. The main difference between this solution and the ideal one is related just to the turn-on of the auxiliary amplifier (Figure 4.35). The behavior of main amplifier drain voltage (Figure 4.34) is more adherent to the theory in the case of class C auxiliary amplifier, since it is constant over the entire auxiliary amplifier range of action. This means that the voltage efficiency of the main amplifier is maximized as desired.

4.3.4 Output Impedance Transformation Network

The impedance transformation network plays an important role in power amplification, since it allows the PA to work with a maximum efficiency. In order to maximize the efficiency, this network must fulfill two targets: provide a precise impedance transformation with reduced losses.

This network can be implemented either outside or inside the chip. The former approach has more degrees of freedom, since the network can be modified until the maximum performance are achieved (eventually with the use of a load pull instrument). But this approach is unsuitable for a large scale realization of a power amplifier, due to the variability of the components used outside the chip and the effect of bondwires (which are integral part of the impedance transformation network). Moreover the need of a balun outside the die (for a differential solution) introduces another element of losses and additional cost. A on-chip solution allows a good control of the impedance transformation supplied and it is more suitable for large scale applications.

There are two possible ways to realize a on-chip impedance transformation network: with an integrated transformer or with a lumped elements network. Both approaches have some advantages and disadvantages. The integrated transformer [22] allows to generate a precise impedance transformation, but the occupied area and the losses introduced can reduce this benefit. In fact there are two possible ways to implement an integrated transformer: with conductors interwound in the same plane or overlaid as stacked metal. The second solution allows to have an higher mutual inductance between the windings of the transformer thus reducing the parasitic capacitance (and so increasing the bandwidth). Since this solution needs two different layers to be integrated, in some technologies where only the top metal has high thickness the losses can be too high.

The proposed solution is a lumped impedance transformation network which also acts as a balun. The ideal implementation of this network is shown in Figure 4.36. This network performs the desired balun operation when driven with a

differential signal, as reported in figure. The differential impedance seen at the resonant frequency ($1/\sqrt{LC}$) at the two sides with a is reported in 4.18 and 4.19.

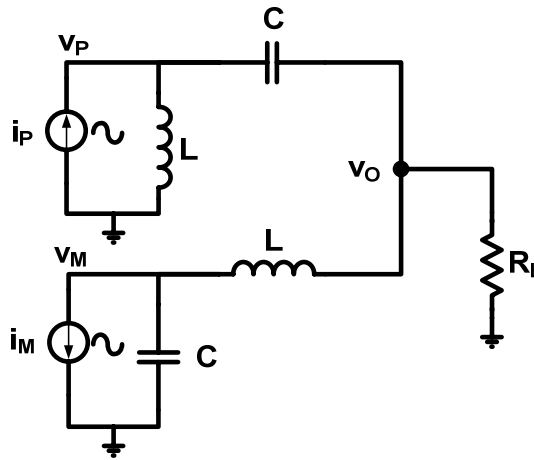


Figure 4.36 Ideal lumped balun

$$\frac{v_P}{i_P} = Z_P = \frac{L}{2CR_L} + j\frac{1}{2}\sqrt{\frac{L}{C}} \quad 4.18$$

$$\frac{v_M}{i_M} = Z_M = \frac{L}{2CR_L} - j\frac{1}{2}\sqrt{\frac{L}{C}} \quad 4.19$$

The residual reactive part can be resonated by a series reactive impedance (capacitive in the upper side and inductive in the lower side). The simulations of this ideal network which provides around 5.5Ω from the 50Ω antenna is shown in Figure 4.37.

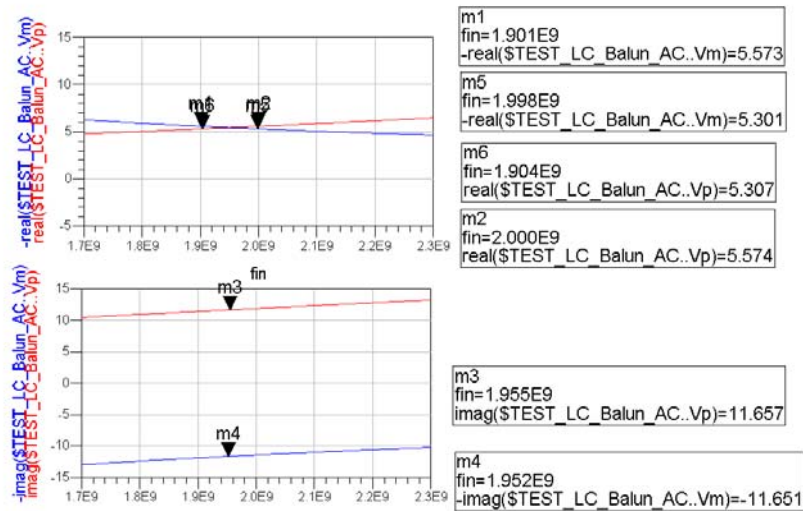


Figure 4.37 Ideal lumped balun simulations

This wideband behavior is unfortunately present only when lossless elements are considered. When a parallel parasitic resistance is considered (in this case 500Ω), the real behavior of this network becomes that of Figure 4.38, where a resonant narrowband behavior is shown, with an abrupt change on the impedance near the resonant frequency (1.95GHz). This behavior is unsuitable for a practical

design, but it can be eliminated by the solution reported in

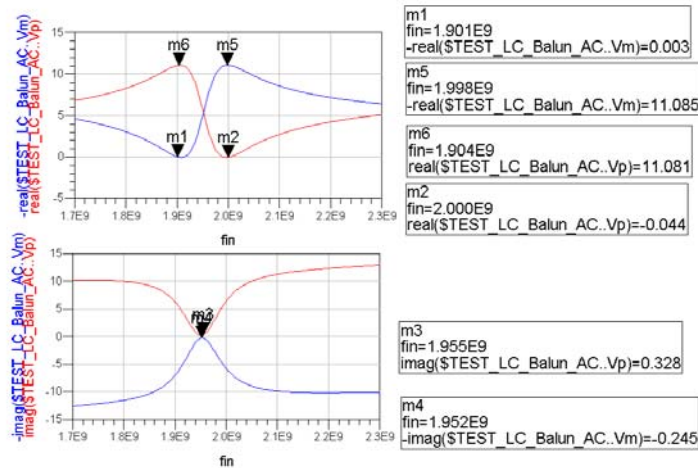


Figure 4.38 Effect of parasitic elements on the ideal lumped balun

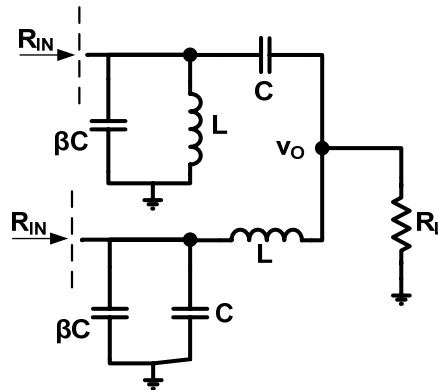


Figure 4.39 Modified ideal lumped balun

The design equations for this solution becomes that reported in 4.20 and 4.21.

$$4.20 \quad R_{IN} = \frac{2LR_L}{4CR_L^2 + \beta^2L} \xrightarrow{\beta \rightarrow 0} R_{IN} = \frac{L}{2CR_L}$$

$$4.21 \quad X_{IN} = \frac{L\beta}{\omega_{RIS}C(4CR_L^2 + \beta^2L)} \xrightarrow{\beta \rightarrow 0} X_{IN} = 0$$

With this modification the resistive and reactive part of the impedance seen at each input have the behavior shown in Figure 4.40. Here it is possible to see that the resistive part of the impedance is almost flat in the range between 1.92GHz and 1.98GHz (the uplink band of interest in the UMTS). Thus a wideband matching is achievable. The reactive part of the impedance is quite low and it should not reduce dramatically the efficiency of the power amplifier.

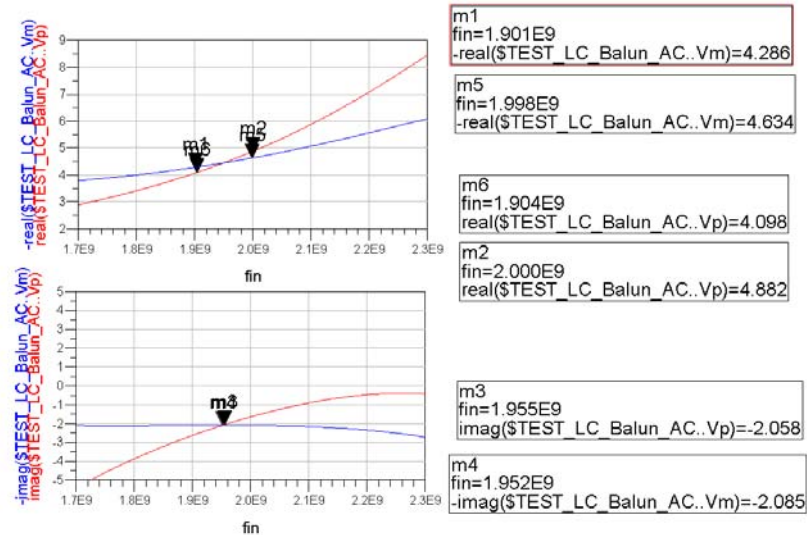


Figure 4.40 Modified lumped balun simulation

The practical realization of this network is depicted in Figure 4.41. The bondwire which connects the common mode pad to the load resistance is resonated by simply using a capacitor outside the chip.

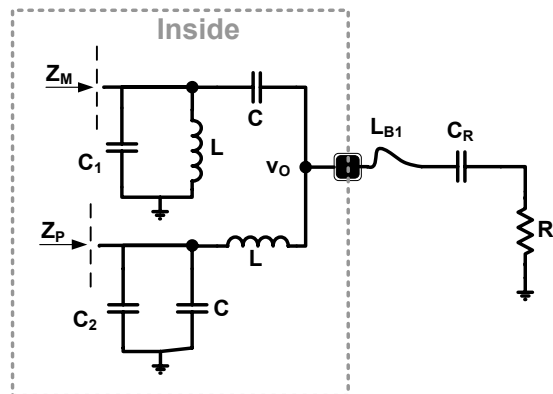


Figure 4.41 Lumped balun implementation

This network has been realized in the CMOS 65nm technology used to design the Doherty power amplifier. It has been tested in a back-to-back configuration apart from the power amplifier, in order to determine the losses introduced by this network. Once the losses are determined it is possible to resize the power amplifier in order to supply the desired power level. The die photograph and the power loss measurement (referred to half of the back-to-back structure) are reported in Figure 4.42 and Figure 4.43.

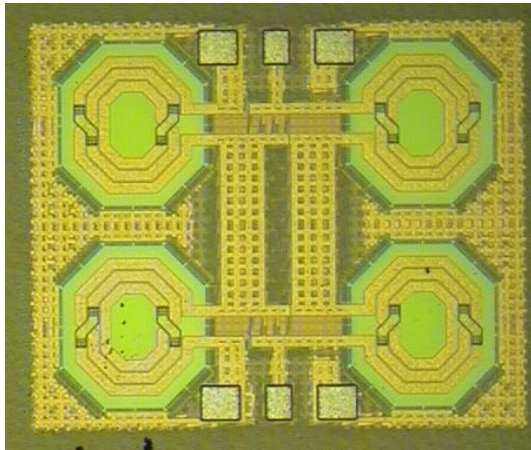


Figure 4.42 Lumped balun test chip

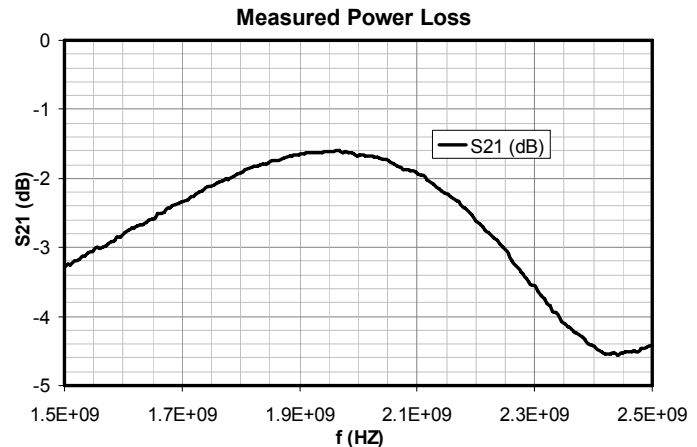


Figure 4.43 Loss measurement of the lumped balun

The power loss introduced by this network at 1.95GHz is around 1.6dB. This loss can be reduced by a fine adjustment of the layout, especially for what concern the lines which connect the two structures. Moreover the loss related to the probe pad are included (but this losses are anyway present in a die, because of the pad for the bondwire).

4.3.5 Pseudo Differential Solution

The previous discussed elements are used in a pseudo differential topology for the Doherty power amplifier in CMOS 65nm. This topology is the best candidate in order to integrate the power amplifier with the rest of the transmitter. In fact the differential topology generates lower spurious current into the substrate, which interfere with the other elements of the transmitter (especially VCOs). The block diagram of this amplifier is shown in Figure 4.44, while the simulations of the overall structure are reported in the next page.

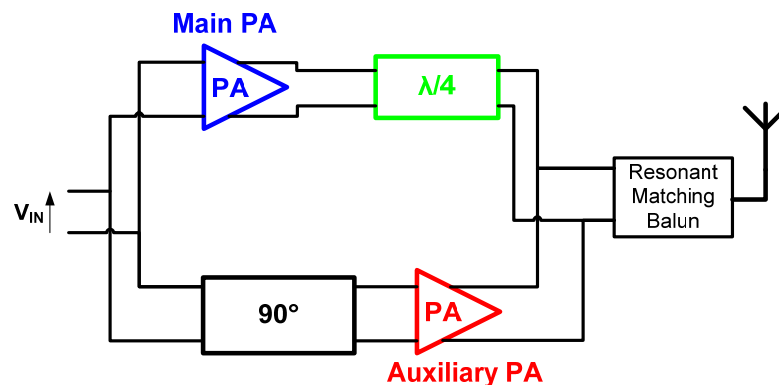


Figure 4.44 Pseudo- differential solution schematic

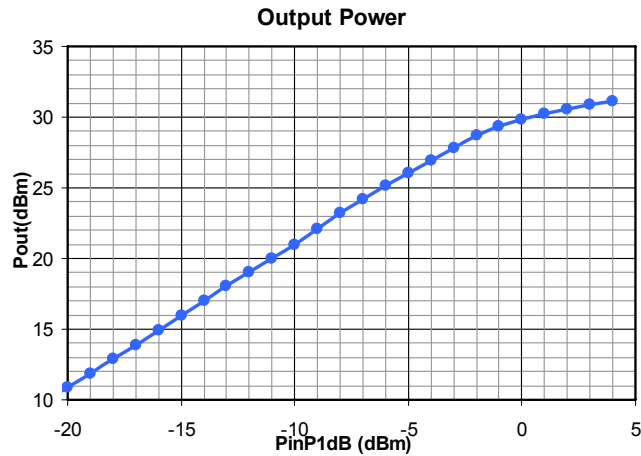


Figure 4.45 Simulated output power of the pseudo-differential solution

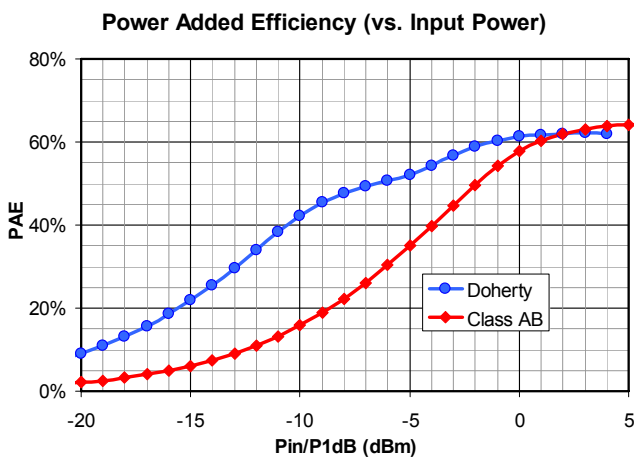


Figure 4.47 Simulated PAE

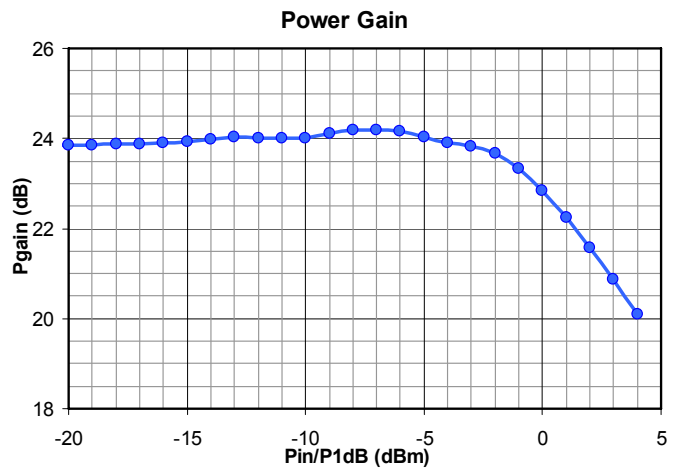


Figure 4.46 Simulated power gain

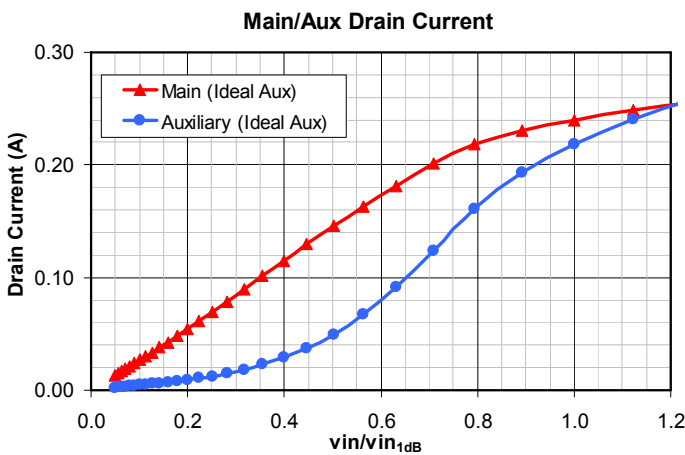


Figure 4.49 Simulated Main/Aux drain current

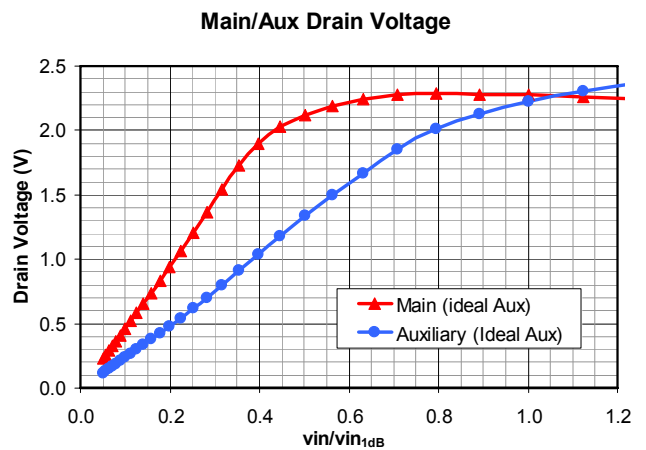


Figure 4.48 Simulated Main/aux drain voltage

This topology is able to deliver an output power at the compression point of 30dBm (Figure 4.45) with an efficiency of 58% (Figure 4.47). The power Gain of the overall amplifier is quite flat (Figure 4.46), showing a small amount of expansion (less than 0.5dB) due to the imperfect matching between the impedances provided at the two outputs by the transformation network shown at

the previous paragraph. Anyway this small power gain expansion doesn't dramatically affect the linearity requirements. The drain voltages and currents (Figure 4.49 and Figure 4.48) recall the behavior of the single end solution.

The efficiency improvement compared to the single stage solution are remarkable, since at 6 dB back off the Doherty amplifier shows an efficiency of 50% against 35% of an ideal class AB solution (which is an overall 40% gain of efficiency). The better performance of the Doherty structure are even more evident at 10dB back off, where its efficiency is more than 40% compared to the 15% of the single stage PA: in practical terms, the duration of the battery used to supply a mobile phone employing this amplifier is nearly doubled compared to a standard solution. Another remarkable result of this topology regards the compact realization of the quarter-wave line, in addition to the optimized (and reduced) area of the main amplifier, which is able to work at its best performance when the auxiliary amplifier is turned off. Another advantage of this solution is that a lumped matching network which works also as a balun allows to limit the elements outside the die, making this amplifier suitable for VLSI applications.

This solution could be further improved by creating a "smart" current shape for the auxiliary amplifier. In fact the dip in the ideal efficiency behavior of the Doherty solution (Figure 2.11) is due to a sub maximal efficiency of the auxiliary amplifier. Supposing to divide the overall auxiliary class C amplifier in several unit element and control them by a DSP, it is theoretically possible to generate any behavior of the drain current. This is obtained by conveniently turn on a specific number of unity elements biased at a higher level (towards the class B) for each input power level, in order to increase the efficiency of the overall amplifier.

4.3.6 Driver stage

The Doherty amplifier needs a convenient driver stage in order to give the necessary input power to the amplifier. Moreover the 90° phase shift at the auxiliary input is needed in order to provide the phase shift which allows the dynamic reduction of the impedance seen at the main amplifier output. This phase shift can be theoretically introduced using the same network which implements the impedance inverter (Figure 4.9). This network would be placed at the auxiliary amplifier input. However this is not the best choice, because of the large signal capacitance variation at the auxiliary input. In fact, as it possible to see in Figure 4.50 where the input capacitance over the normalized input power is reported, the input capacitance has a significant variation after the breakpoint. This means that the overall resonant frequency and phase shift introduced by the quarter-wave line (which will be in parallel to the input capacitance) will be dependent on the amplitude of the input signal, causing a wrong behavior of the Doherty structure.

Auxiliary input capacitance

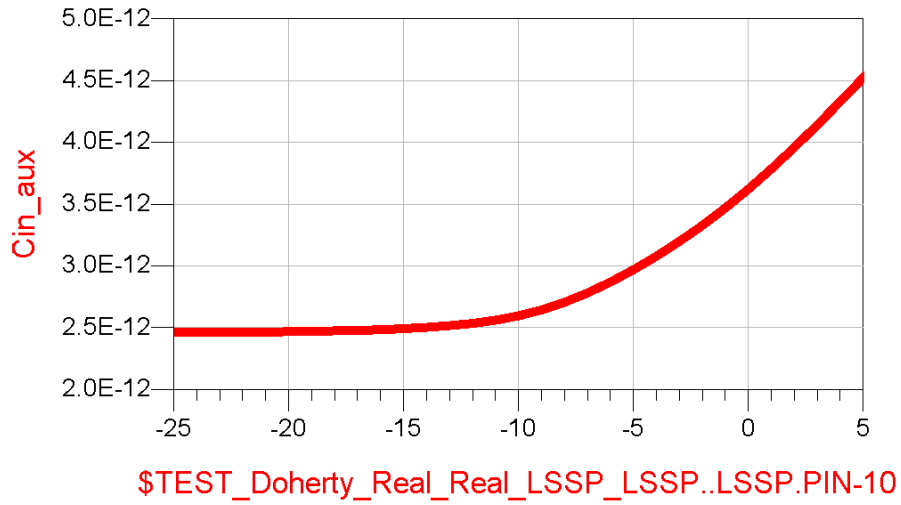


Figure 4.50 Auxiliary input capacitance variation for large signals

This behavior encourages the implementation of a system able to generate a desired phase between the main and the auxiliary amplifier. The proposed system is shown in Figure 4.51.

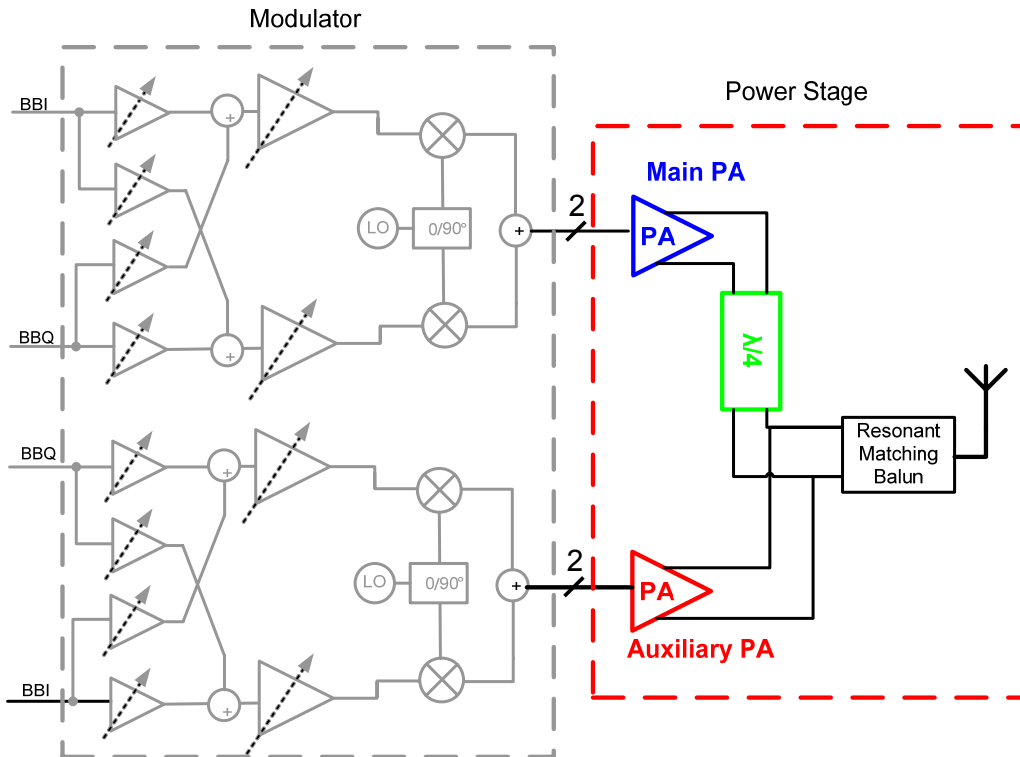


Figure 4.51 Complete transmitter with arbitrary phase shift

Each side is able to provide a phase shift which can be digitally controlled at the baseband with a gain variation of the amplifiers in the modulator. The overall structure is differential.

The working principle of this modulator can be explained referring to Figure

4.52. The C_1 and C_2 signal are obtained by an amplified weighted sum of the A_1+B_2 and A_2+B_1 signals, which are the phase a quadrature components of the modulated signal. The phase of the modulated signal can be changed by a convenient amplification of the A_1 , A_2 , B_1 , B_2 signals. The signal is then upconverted. If two of these structure are used to feed the input signal to the main and the auxiliary amplifier, it is possible to generate an arbitrary phase shift between the two input signals, allowing to set the phase variation at different input powers in order to maximize the linearity. This can be performed by a calibration algorithm, which can be digitally implemented.

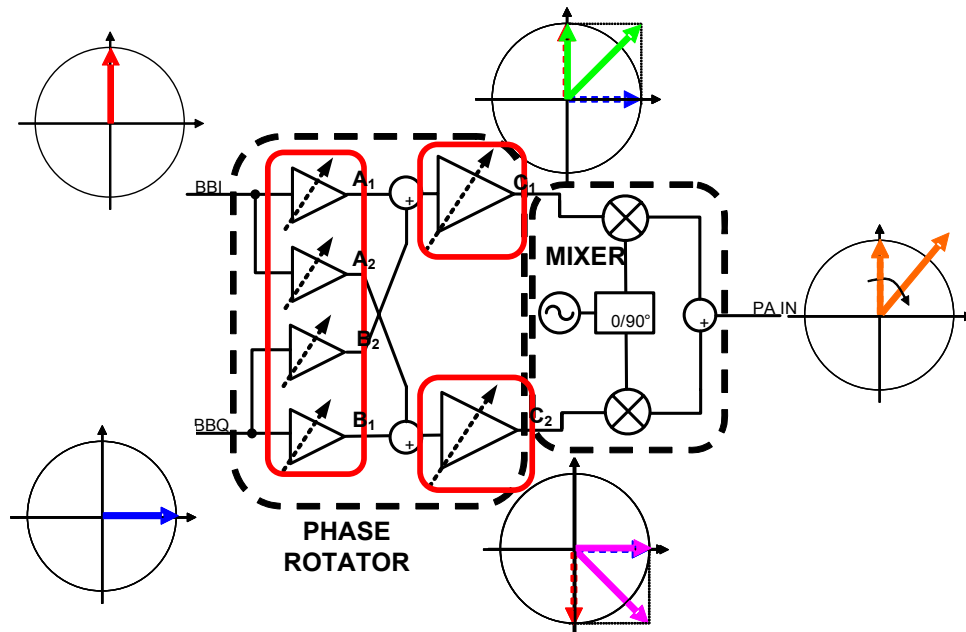


Figure 4.52 Phase rotator and up-conversion

This system allows to realize a fully integrated transmitter comprehensive of the power amplifier. This would be a great achievement since the most challenging accomplishment in RF transmitters is to integrate the power amplifier with the overall transmitter. This thesis concludes with the design of the power stage of this structure, while the overall realization is still object of further research.

4.4 Conclusion

The design of an integrated Doherty Power Amplifier in CMOS 65nm technology has been addressed. This efficiency enhancement technique allows to increase the efficiency of a linear power amplifier over a wider range if compared to the standard class AB implementation. A class C auxiliary amplifier allows to maintain the efficiency close to the behavior obtained with an ideal auxiliary amplifier with an abrupt turn-on. An integrated lumped balun has been designed and tested in order to implement a on-chip impedance transformation. A complete integrated transmitter employing this power amplifier has been introduced and its working principle explained.

References

- [18] Steve C. Cripps, “Advanced Techniques in RF Power Amplifier Design”, *Artech, 2002*
- [19] Hammi, O.; Sirois, J.; Boumaiza, S.; Ghannouchi, F.M.; “*Design and performance analysis of mismatched Doherty amplifiers using an accurate load-pull-based model*” *Microwave Theory and Techniques, IEEE Transactions on* Volume 54, Issue 8, Aug. 2006 Page(s):3246 - 3254
- [20] Elmala, M.; Paramesh, J.; Soumyanath, K.; “*A 90-nm CMOS Doherty power amplifier with minimum AM-PM distortion*” *Solid-State Circuits, IEEE Journal of* Volume 41, Issue 6, June 2006 Page(s):1323 - 1332
- [21] Iwamoto, M.; Williams, A.; Pin-Fan Chen; Metzger, A.G.; Larson, L.E.; Asbeck, P.M.; “*An extended Doherty amplifier with high efficiency over a wide power range*” *Microwave Theory and Techniques, IEEE Transactions on* Volume 49, Issue 12, Dec. 2001 Page(s):2472 - 2479
- [22] Long, J.R.; “*Monolithic transformers for silicon RF IC design*” *Solid-State Circuits, IEEE Journal of* Volume 35, Issue 9, Sept. 2000 Page(s):1368 - 1382

Appendix

Thermal Effects

Temperature variations affect the performance of a power amplifier in terms of gain and maximum output power reduction. Before looking at the electrical effects determined by a temperature variation in the silicon die, a brief introduction at the concept of equivalent thermal circuit will be given.

A.1 Equivalent thermal circuit

According to the first law of thermodynamics, “The change in internal energy of a system is equal to the heat added to the system minus the work done by the system”. From the PAs point of view, the power amplification of an electronic signal is obtained by converting (work) the DC supply power in RF power (energy). The part of the DC supply that is not converted into RF power becomes heat. Thus, the quantity of heat generated by the Power Amplifier depends by its efficiency.

The heat generated by the amplification mechanism must be conveniently dissipated by a heat sinker, in order to minimize the temperature variation into the die. This temperature variation depends by the characteristics of the heat sinker and how the die is placed in contact with it. It can be predicted and modeled by the use of an equivalent thermal circuit, which links the electrical and the thermal environment thanks to the concept of thermal resistance.

The thermal resistance is a parameter which allows to predict the temperature variation of a given element when an electrical power passes through it. The thermal resistance depends by its thermal conductivity (σ_{th}) and its area (A) and thickness (t):

$$A.1 \quad R_{th} = \frac{1}{\sigma_{th}} \frac{t}{A} \left[\frac{^{\circ}C}{W} \right]$$

For example, when a power of 1W is dissipated by an element with a thermal resistance of 1 °C/W, its temperature increases of 1°C respect to the environment temperature. The thermal conductivity is the property of a material that indicates its ability to conduct heat. Typical units are SI: W/(m·K) and English units: Btu/(hr·ft·°F) (to convert between the two, use the relation 1 Btu/(hr·ft·°F) = 1.730735 W/(m·K)).

Once defined the thermal resistance it is possible to draw the equivalent thermal circuit of an electronic system, in order to calculate the temperature variation of the circuit. The thermal circuit can be solved by using KCL and KVL and a modification of the first Ohm’s law:

$$A.2 \quad \Delta T = P \cdot R_{th}$$

Where ΔT is the temperature variation of an element with thermal resistance R_{th} when a

DC power P flows into it. It is evident that the DC power can be modelled by an analogy to an electrical circuit as a current and temperature can be modelled as a voltage (see Figure A.1).

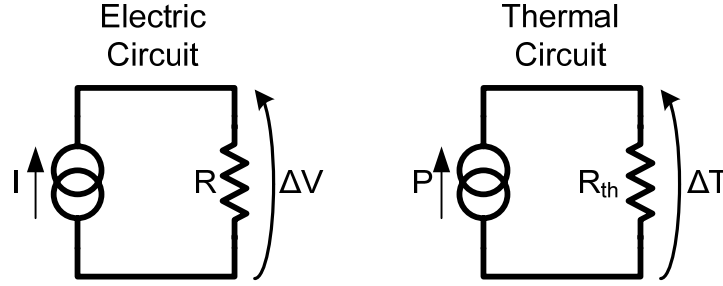


Figure A.1 Electric circuit compared to thermal circuit

Generally the reference is the ambient temperature, and the variations in the circuit are calculated in respect to it.

As an example we can calculate the temperature variation of a power amplifier. Suppose to have an ideal class B PA that delivers a power of 1W with an efficiency of 78.5%. Since the efficiency is the ratio between the RF power (P_{RF}) delivered to the antenna and that one provided by the power supply (P_{DC}), in these conditions we have:

$$A.3 \quad P_{DC} = \frac{P_{RF}}{\eta} = 1.27W$$

This means that 270mW are converted in form of heat (since 1W is converted in RF power and delivered to the load). If the power amplifier is glued to an heat sinker with a thermal resistance of 40°C/W, and if we consider negligible the die's and glue's thermal resistance, the temperature variation of the heat sinker (and thus the temperature of the die) will be:

$$A.4 \quad \Delta T = P \cdot R_{th} = 11^\circ C$$

For a class B PA, this condition happens only at the maximum deliverable RF output power, when the efficiency reaches its maximum value. But in the real applications (like UMTS) the signal has a variable amplitude. This means that the average transmitted power can be significantly lower compared to the maximum power. Since the efficiency drops when the output power decreases, the temperature variation will be consequently different respect to the maximum power condition. In order to evaluate the temperature in back-off condition, now suppose to transmit half of the maximum power (0.5W). For a class B ideal PA, the efficiency will be around 55%. The DC power will be:

$$A.5 \quad P_{DC} \Big|_{@ \frac{P_{MAX}}{2}} = \frac{0.5}{0.55} = 910mW$$

In this case the fraction of the DC power that becomes heat is 410mW (which is larger than the previous case). Therefore, the temperature variation respect to the ambient will be:

$$A.6 \quad \Delta T \Big|_{@ \frac{P_{MAX}}{2}} = P \Big|_{@ \frac{P_{MAX}}{2}} \cdot R_{th} = 16.4^\circ C$$

This last result shows that the efficiency impacts the temperature variation more than the transmitted power. It is then necessary to maximize the efficiency at the average output power for a given thermal circuit in order to minimize temperature variations.

A.2 Electrical Effects caused by temperature

Temperature variations have an influence on the electrical behavior of a power amplifier affecting power gain, linearity, maximum output power and efficiency.

The DC power of a linear PA varies dynamically with the amplitude of the input signal. This causes a temperature variation. As shown in the previous paragraph, temperature can have significant variations for different output levels. This changes the performance of the active and passive devices. For what concern the former, temperature affects the small signal gain and the substrate resistance, while in the latter (especially inductors) an increased temperature gives a reduction of the quality factor. Thus, a dynamic variation of the temperature cause a dynamic gain variation affecting the linearity performance.

The effect of the temperature on the power gain also depends how the transistor is biased. If it is biased with a fixed voltage, a temperature variation gives a quiescent current variation changing the working class and the gain of the power amplifier. On the contrary, if it is biased with a fixed current, the temperature variation does not change the power gain.

The maximum output power depends on the maximum output current that the transistor can supply. When the temperature increases, the maximum current deliverable by the transistor decreases. Since the output impedance and the supply voltage do not change, this determines a reduction in the maximum output power and efficiency.

As a practical example we can look at the simulations of a bipolar class AB power amplifier in a common emitter topology. Let's first consider a constant-voltage bias.

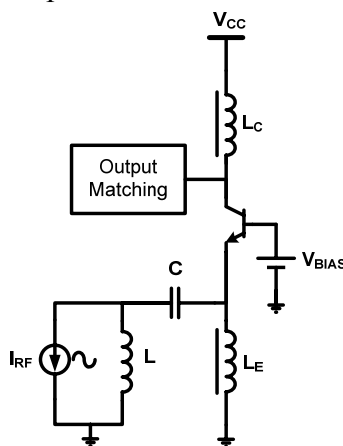


Figure A.2 Constant voltage bias

In this case the common-base stage is biased with an ideal voltage source. This one fixes the voltage at the transistor's base which remains constant during any temperature variation.

The I_{RF} sinusoidal current generator replaces the driver stage before the power amplifier, which is coupled with an LC resonant network.

The output matching network transforms the 50Ω antenna impedance in a lower impedance, so that the transistor makes a power amplification of the input signal.

In the next figures the Output Power, Gain and Efficiency are drawn for this circuit under different temperature conditions.

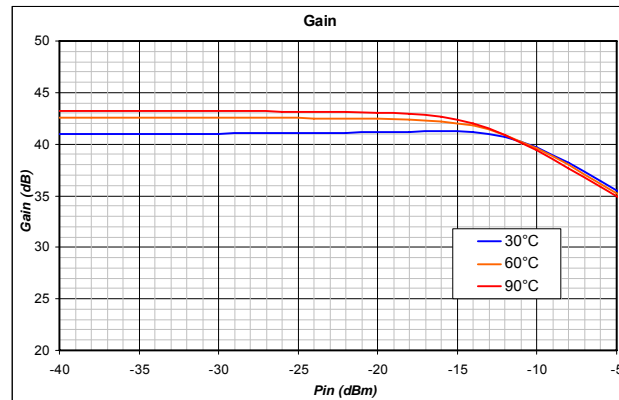


Figure A.3 Power gain with constant voltage bias (simulations)

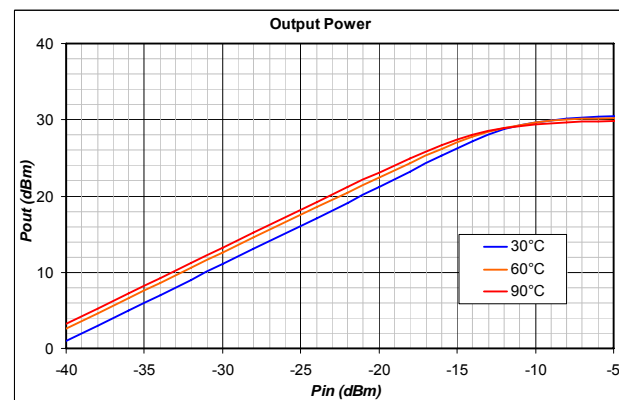


Figure A.4 Output power with constant voltage bias (simulations)

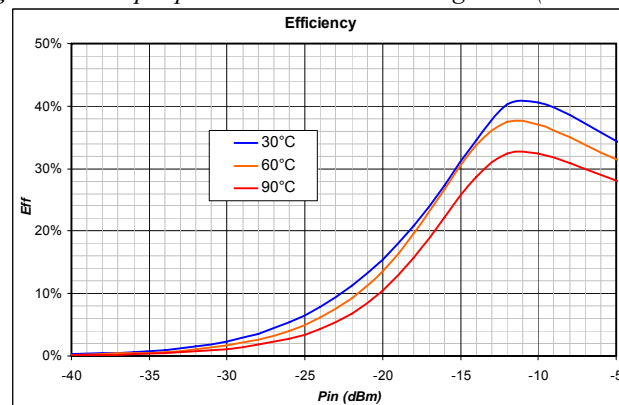


Figure A.5 Efficiency with constant voltage bias (simulations)

As we can see looking at the previous graphs, the power gain changes for different temperatures. This means that if the input signal variation causes a significant temperature variation, the compressive or expansive behavior of the power gain will be temperature dependent. In other words, at each temperature the punctual gain value for a given input power lies on the power gain curve for that temperature.

Fortunately, a modulated signal for a 3G system does not cause a substantial temperature variation around the average transmitted power. The dissipated power for a given average output power must be taken into account in order to predict the circuit temperature and then assume it as a constant temperature for the system. Also the ambient temperature variation must be taken into account in order to predict the circuit performance in all the possible cases.

The efficiency is also affected by temperature in such a bias method, where the efficiency decreases with a temperature increment. The design of a power amplifier must consider the effective conditions where the circuit is going to work, in order to take the convenient countermeasure to reach the design goals. This means that the testing board thermal circuit must be considered as a design constraint.

A different bias approach is the constant-current bias, showed in Figure A.6.

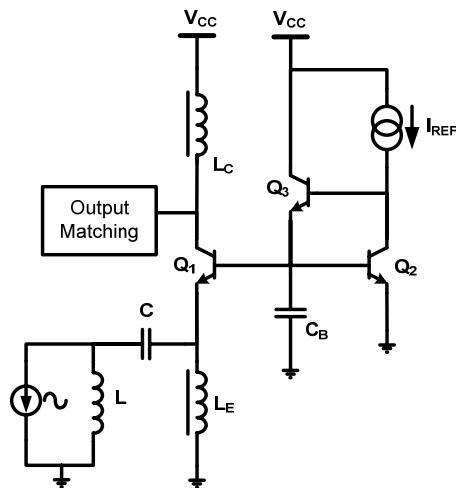


Figure A.6 Constant current bias

The reference bias current I_{REF} is mirrored and scaled in Q_1 by Q_2 . The Q_3 transistor works as a buffer, in order to provide the necessary base current at the high signal levels.

This bias topology keeps the DC current constant over temperature, while the V_{BE} of Q_1 varies accordingly with it. Moreover, a low base impedance at DC and all of the harmonics of the signal is provided in order to avoid breakdown issues.

The next pictures show the performance of this bias topology when a RF signal is applied over various temperatures.

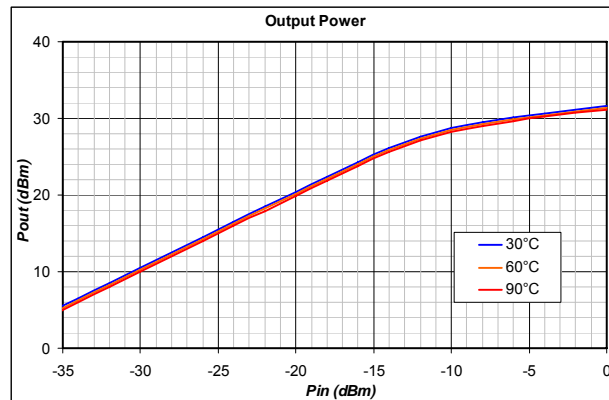


Figure A.7 Output Power with constant current bias (simulations)

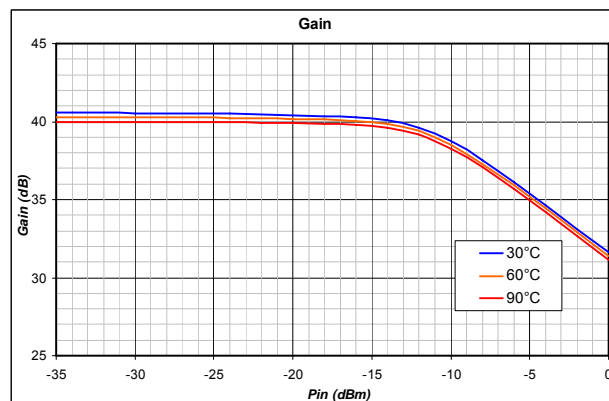


Figure A.8 Power Gain with constant current bias (simulations)

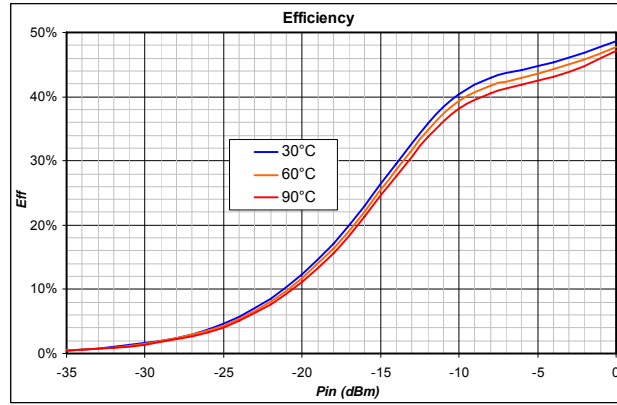


Figure A.9 Drain Efficiency with constant current bias (simulations)

In this case the gain and efficiency variation due to temperature are smaller compared to the constant voltage bias. This bias topology seems to be more appealing compared to the previous one. Unfortunately it has the drawback of the real implementation of the current source which generates I_{REF} . In fact, if this current generator is implemented by a simple off-chip resistor, the bias current will be temperature dependent, since the voltage V_{C2} changes with temperature and so I_{REF} changes. This effect will be discussed on a real implementation in the next paragraph.

Another bias approach is the constant-Gm bias, showed in Figure A.10.

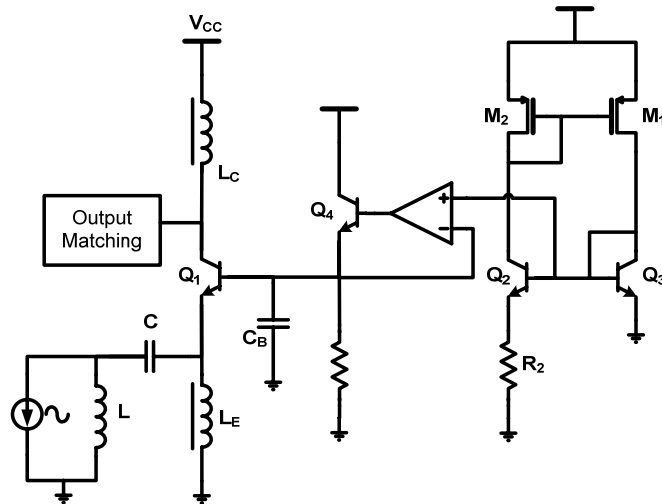


Figure A.10

In this circuit the DC current flowing in Q_2 is mirrored in Q_1 . The DC current I_{REF} depends by the scaling factor k between Q_2 and Q_3 and the resistance R_2 with the relation (M_1 and M_2 have the same size):

$$A.7 \quad I_{REF} = V_T \frac{\ln k}{R_2}$$

And the small signal gain of Q_2 is:

$$A.8 \quad g_{m_2} = \frac{I_{REF}}{V_T} = \frac{\ln k}{R_2}$$

which is insensitive from temperature.

The small signal gain of Q_1 has the same property of g_{m_2} , that is why this solution is called constant-gm bias. The buffer made by Q_4 and the OP-AMP provides low impedance and base current at the highest signal levels.

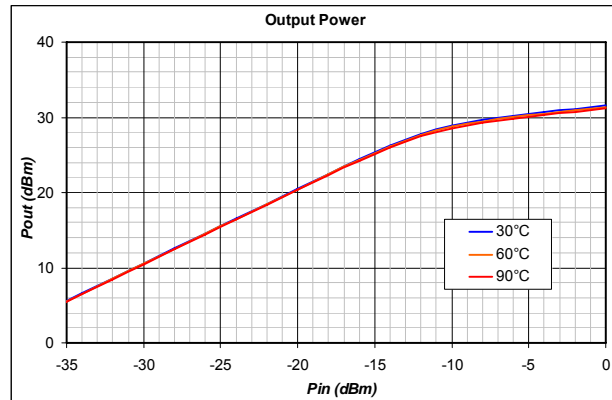


Figure A. 11 Output Power with constant gm bias (simulation)

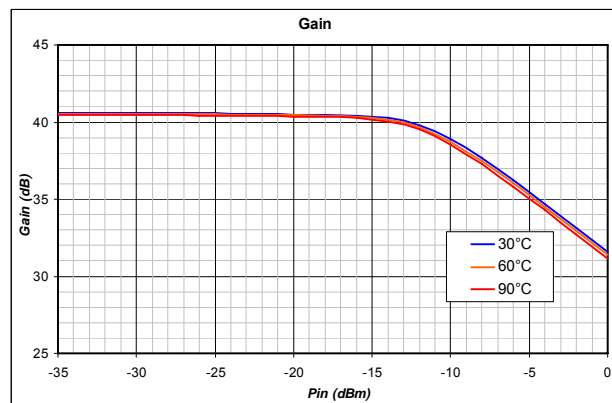


Figure A. 12 Power Gain with constant gm bias (simulation)

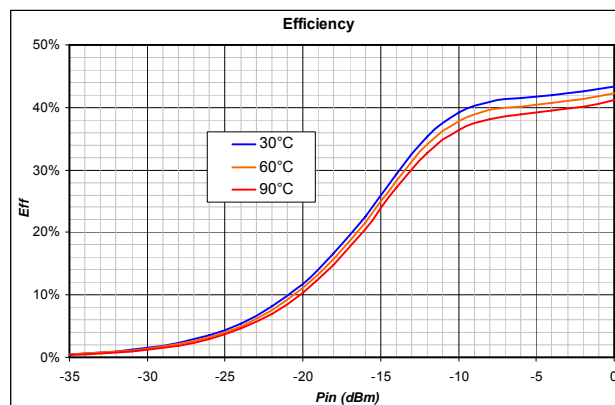


Figure A. 13 Efficiency with constant gm bias (simulation)

This solution seems to have the same performance as the constant current bias, since the output power, gain and efficiency doesn't change significantly for different temperatures. Generally speaking, the choice between these two solutions depends from the specific case in which they are going to be used.

A.3 Constant current bias on a real circuit

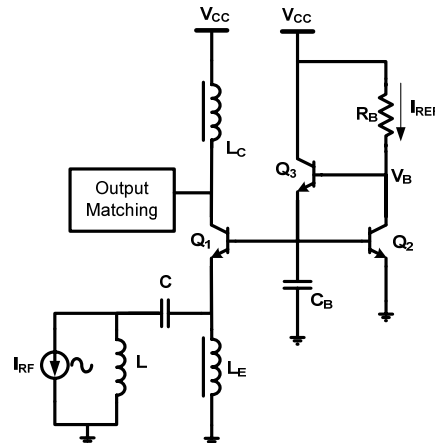


Figure A.14 Constant current bias - implementation

A common base linear PA has been design and tested, using a constant current bias topology. In this case, the reference current generator has been replaced by a simple resistor, as shown in Figure A.14.

The reference current depends by the difference between V_{CC} and V_B . Since the latter depends on temperature, also the reference current will be temperature dependent: this means that the temperature variation will impact the circuit performance more than in the ideal case.

The entire pseudo-differential circuit including the driver stage is shown in the next figure.

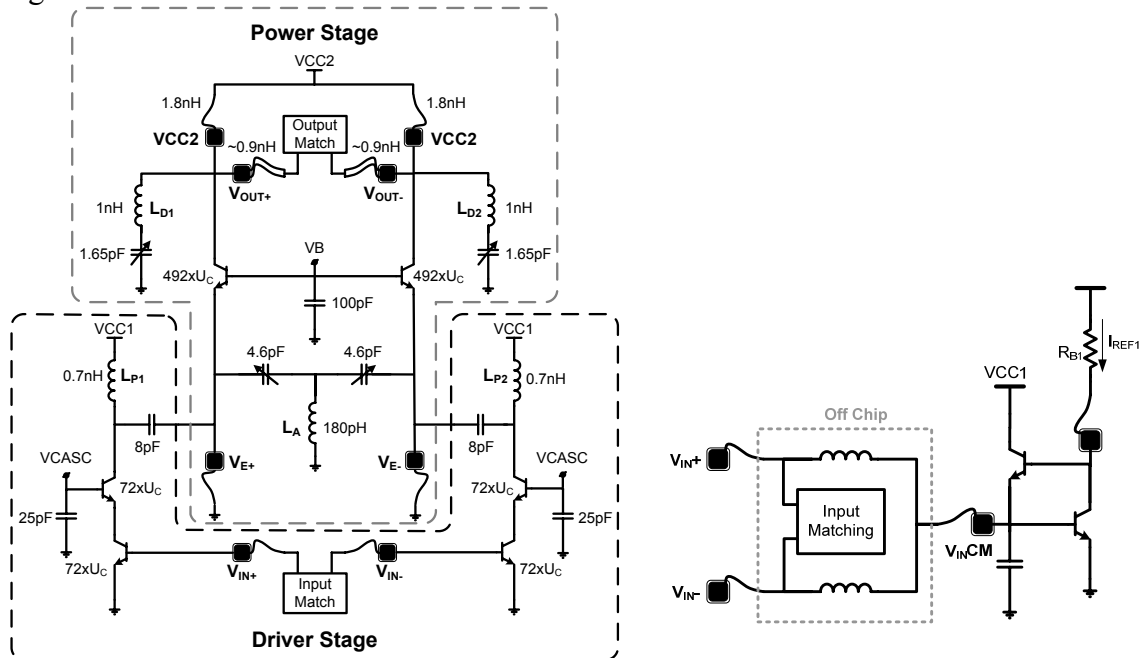


Figure A.15 Complete pseudo-differential schematic

Since the bias circuit is now sensitive to temperature, a good thermal circuit must be used in order to keep temperature and its variations as low as possible. The thermal circuit is made up of the silicon die and the case where it is bound (e.g. the testing board or a heat sinker).

Three different thermal circuit has been investigated and measured, and their effect on the circuit performance has been simulated. In the first one, the die was simply bound to the FR4 test board, without any heat sinker.

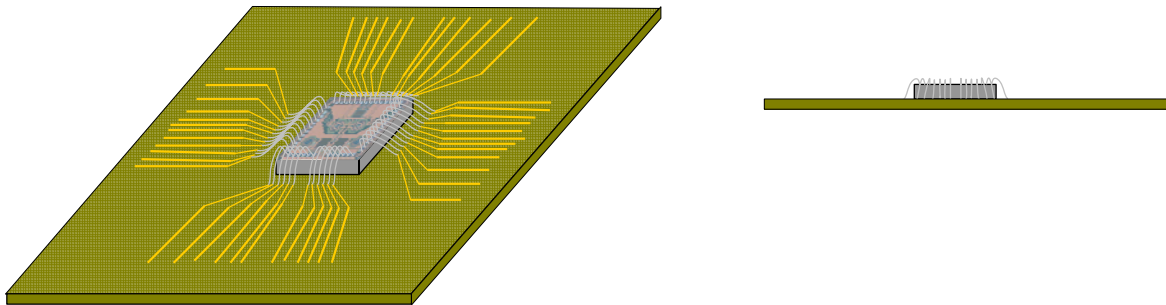


Figure A.16 *Thermal circuit without heat sinker*

In the second case, the die was bound on a $30^{\circ}\text{C}/\text{W}$ heat sinker, while in the third case a $10^{\circ}\text{C}/\text{W}$ heat sinker was used.

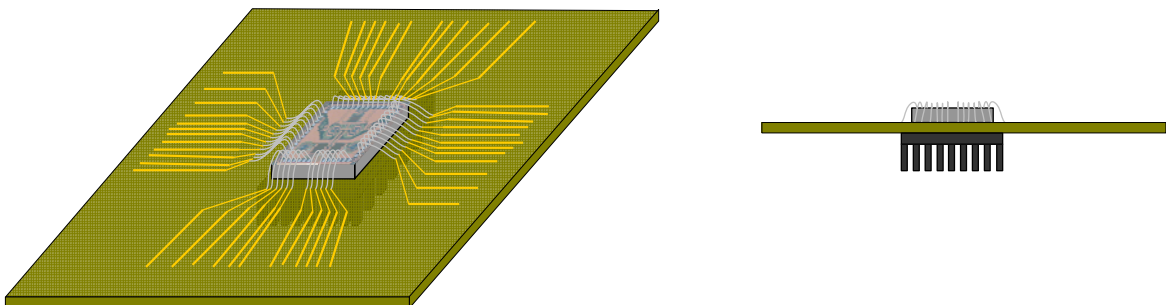


Figure A.17 *Thermal circuit with $25^{\circ}\text{C}/\text{W}$ heat sinker*

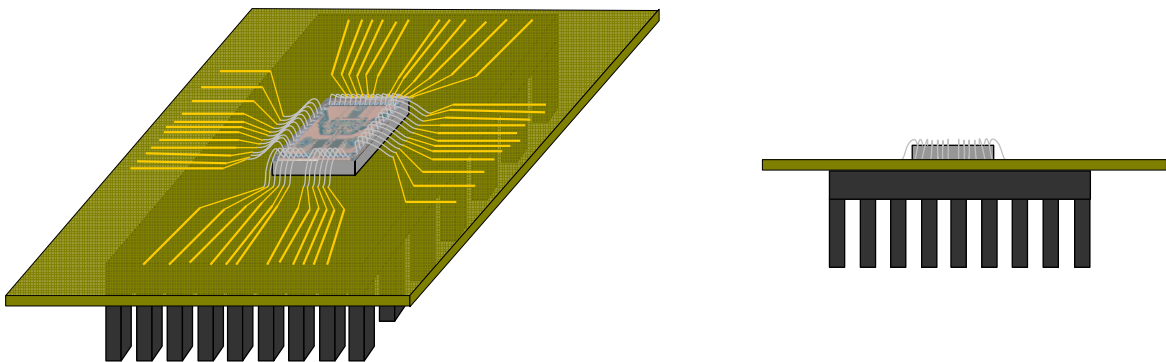


Figure A.18 *Thermal circuit with $10^{\circ}\text{C}/\text{W}$ heat sinker*

In the first case a common glue ha been used, without knowing its thermal characteristics. In the other cases a particular glue with good thermal characteristic has been used to bind the circuit to the heat sinker, so its effect on the thermal circuit can be considered negligible. Also the die's thermal resistance can be considered negligible. Since the silicon has a thermal conductivity of $130 \text{ W}/^{\circ}\text{C}\cdot\text{m}$ and the die has a surface of 2.76 mm^2 and a thickness of 0.8 mm , its thermal resistance is around $2.22 \text{ }^{\circ}\text{C}/\text{W}$.

The measurement where performed by measuring the V_{INCM} voltage variation for different DC bias conditions i.e. for different power dissipations. Due to the thermal resistance, this voltage changes when the dissipated power changes: this means that also temperature is changing. Each measurement point has been compared with a DC

simulation, where the simulation temperature were iteratively changed in order to fit the V_{INCM} measured voltage with the simulated one. With this indirect measurement it is possible to calculate the thermal resistance of the circuit. This estimation is reported in Figure A.19:

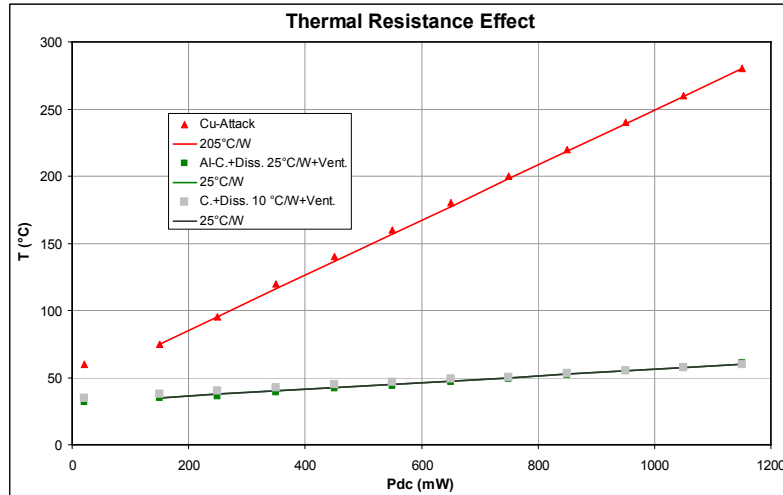


Figure A.19 Simulation and measurement of the thermal circuits

The measurement shows that the first configuration has a thermal resistance of around 205 °C/W which is very high and it is not suitable for a PA application. In the other two cases the thermal resistance is around 25°C/W although the two heat sinkers have different thermal resistance. This behavior is due do the very high area of the wider sinker (which has the lower thermal resistance). In this case the die can be considered as a punctual heat source making a not-uniform heating of the heat sinker, resulting in a bigger thermal resistance.

The thermal resistance estimation of the configurations of Figure A.16 and Figure A.17 has been validated by a thermal image of the die when dissipating around 0.8 W. These measurement show a good agreement with the estimation performed with the simulations.

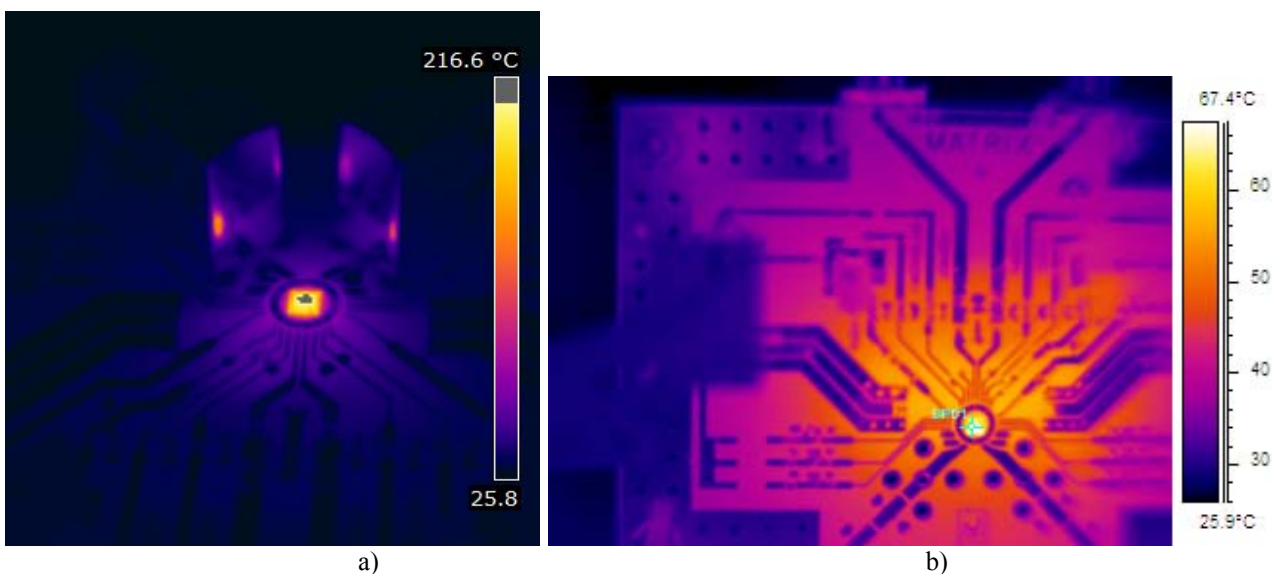


Figure A. 20 Thermal image of the circuit without heat sinker a) and with a 25 °C/W heat sinker

