Investigation of 16 nm FinFET and 22 nm FD-SOI CMOS Technologies for Millimeter-Wave Power Amplifiers

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für
Elektrotechnik und Informationstechnik
des Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

M.Sc. Mario Lauritano
geb. in Palermo, Italien

Tag der mündlichen Prüfung: 20.09.2024

Hauptreferent: Prof. Dr.-Ing. Ahmet Çağri Ulusoy
Korreferent: Prof. Dr.-Ing. Vadim Issakov

Acknowledgements

This thesis is the result of over five years of intense work in the Analog and RF Enablement (ARE) team at Intel Germany in cooperation with the Insitute for High Frequency Techniques and Electronics (IHE) at the Karlsruhe Institute of Technology (KIT). Doing research in an industrial environment is a great opportunity but it comes with the challenge of identifying topics of mutual interest for both industry and academia. Furthermore scientific work, by its very nature, is full of uncertainties and can often require changes in perspective and direction, and in this regard my doctorate was no exception. For these reasons my time as PhD researcher has been characterized not only by intense and continuous learning, but also by tough challenges, which every now and then made me doubt wether I would reach the end. Therefore, I would like to express my gratitude to all the people who, in one way or another, supported me in addressing and overcoming these challenges.

First and foremost, I would like to thank my advisor at Intel, Dr. Peter Baumgartner, for his tireless commitment to my work, his inspiring understanding of the subject matter and his incomparable problem-solving attitude. I am deeply greatful to Prof. Dr. Jasmin Aghassi-Hagmann for supervising the first part of my PhD with consistently high motivation and to Prof. Dr.-Ing. Ahmet Çağri Ulusoy for accepting to become my academic advisor in 2021. His strong impact on the focus of the thesis and his constant support have been invaluable. I would like to extend my gratitude to Dr. Gerhard Knoblinger from Intel Austria for helping me find the perfect team within the company for my doctoral research. I also wish to thank my fellow current and former PhD students at Intel Germany, Dr. Richard Hudeczek, Carla Moran Guizan, Felix Last, Lukas Schramm, Puneet Singh and Emanuele Groppo. We have shared a great deal of technical discussions and learned together day by day how to be good engineers and researchers.

I am very grateful to all the members of the ARE team for being not only excellent engineers but also great people. I would like to acknowledge Dr. Michael Langenbuch and Dr. Daniel Sira for their guidance during the early phase of

my PhD, Hui Zhang, Dr. Alexander Bechtold and Manir Mohammad for their support in the lab, and Lukas Epple, Jean-Paul Dochoyan and Ali Bitar for their outstanding help with the simulation tools. Special thanks also to Dr. Richard Geiger for being both a professional role model and a very good friend, and to Sara Schlüter for the many enjoyable coffee breaks and the fun conversation. Outside the ARE team, I would like to acknowledge Dr. Saeid Daneshgar, Dr. Steven Callender, Dr. Run Levinger, Dr. Fabian Cossoy and Dr. Ritesh Bhat for their valuable insights on circuit design, and Luis Abreu and Clément Mélen for their support with reliability topics. I also express my sincere appreciation to Prof. Ulusoy's PhD students at KIT, above all Alexander Haag, Tsung-Ching Tsai and Tai-Yu Kuo.

Lastly, I would like to mention the exceptional contribution of the people who accompanied me in my personal life throughout this journey. First and foremost, my mother Valeria Adamo and my sister Gloria Lauritano, two great women who have always been un unwavering source of security. My life-long friend Stefania Di Bella for our frequent phone conversations and get-togethers, which never failed to cheer me up, even in the darkest moments. My friends in Munich, particularly Giulia Vichi, Ana Becker, Daniela Cardoso, Mustafa Gaja, Alexandra Kegai, Pedro Ribeiro and Jurij Snigirjov, for being with me all this time and making me feel at home. Thanks a lot also to my college friends from Turin, Edoardo Paganoni, Marco Cipolato, Umberto Fugiglando and Francesca Mosca, and those from my year abroad in Chicago, Fabio Ghiozzi, Stefano Arseni and Sara Cantoni. I am deeply grateful that we have remained close despite living in different places for so many years. I would also like to mention Antonio Galia, Roberta Migliorino, Giusi Di Bella, Alessandro Morabito, Lorenzo Barbaro, Marialaura Rosso, Federica Bustinto, Leonardo Arpino, Luigi Marzocchella and Fabrizio Rinaldi for being, each in their own way, excellent friends. Finally, a heartfelt thanks to the fantastic people I met during my backpacking trip to South America, especially Zuli Tatiana Perez, Hanna Sutterer, Laura Daza and Daniel Cañas, for being with me during the best time of my life.

Zusammenfassung

Moderne drahtlose Transceiver verwenden digital-intensive Schaltungsarchitekturen, um die Vorteile von CMOS-Technologien zu nutzen, erfordern jedoch immer noch analoge Frontend-Komponenten bei Mikrowellenfrequenzen, um Signale über die Luft zu senden und zu empfangen. Die begrenzte Transistorgeschwindigkeit und Durchbruchspannung von CMOS-Technologien stellen erhebliche Hürden für die Entwicklung von Mikrowellenschaltungen dar, insbesondere von Leistungsverstärkern. Aus diesem Grund wurde das analoge Frontend üblicherweise in einem separaten integrierten Schaltkreis auf Basis von leistungsstarken, aber hochpreisigen III-V-Halbleitersubstraten implementiert. Die fortlaufend steigenden Betriebsfrequenzen, die inzwischen in den Millimeterwellenbereich reichen, haben zu hohen Anforderungen an die Transistorgeschwindigkeit, aber auch zu einer erhöhten Realisierbarkeit von phasengesteuerten Antennenarrays geführt. Dies hat CMOS zu einer möglichen Alternative zu III-V-Technologien gemacht, wodurch die Integration eines gesamten Transceivers auf einem einzigen Chip ermöglicht wurde. Obwohl die Geschwindigkeit von CMOS-Transistoren mit der Technologieverkleinerung einen positiven Trend gezeigt hat, hat die Spannungsfestigkeit kontinuierlich abgenommen, was die Entwicklung von Leistungsverstärkern in fortgeschrittenen CMOS-Knoten zunehmend herausfordernd macht. Die Skalierung jenseits des 28-nm-Knotens wurde durch die Einführung von FinFET- und SOI-Prozessen ermöglicht. Die überlegene digitale Leistung von FinFET Technologie und ihr besseres Skalierungsverhalten haben es zur bevorzugten Wahl für digitale Anwendungen gemacht, während sich SOI als besser geeignet für analoge Schaltungen erwiesen hat. Dies hat die Frage aufgeworfen, ob es vorteilhafter ist, das gesamte System auf einem FinFET-Chip zu integrieren oder die digitalen und analogen Funktionen zwischen FinFET und SOI aufzuteilen. Diese Arbeit zielt darauf ab, die Leistung der 16-nm-FinFET- und 22-nm-FD-SOI-Prozesse für die Entwicklung von Millimeterwellen-Leistungsverstärkern zu untersuchen und zu vergleichen. Sie behandelt Aspekte der Schaltungsentwicklung sowie der Bauelementmodellierung, wobei ein besonderer Schwer-

punkt auf der Charakterisierung von einzelnen Bauelementen wie Transistoren und Transformatoren liegt. Die Untersuchungen werden im E-Band (60-90 GHz) durchgeführt, einem Frequenzbereich von Interesse für verschiedene Anwendungen, vor allem für Automotive Radar. Im Gegensatz zu den meisten bestehenden Veröffentlichungen zu diesem Thema liegt der Fokus dieser Arbeit darauf, den Einfluss von Technologieeigenschaften auf die Schaltungsperformance zu verstehen, anstatt innovative Schaltungsarchitekturen für eine bestimmte Technologie vorzuschlagen. Die Arbeit ist wie folgt aufgebaut: Kapitel 1 gibt eine Einführung in millimeterwellenbasierte drahtlose Verbindungen und einen Überblick über die relevanten Anwendungen. Kapitel 2 behandelt die Grundlagen von Millimeterwellen-Leistungsverstärkern und der dafür benötigten Halbleitertechnologien, wobei der Schwerpunkt auf stark miniaturisierten CMOS-Prozessen liegt. Kapitel 3 schlägt ein neues Konzept für die Messung des Gate-Widerstands vor, das anschließend mithilfe von Teststrukturen in der 16-nm-FinFET Technologie validiert wird. Kapitel 4 bis Kapitel 6 umfassen den Kern dieser Arbeit, der in der Entwicklung einer algorithmischen Methodik für das Design von Millimeterwellen-Leistungsverstärkern besteht, um einen fairen Vergleich zwischen verschiedenen Prozessen zu ermöglichen, in diesem Fall 16-nm-FinFET und 22-nm-FD-SOI. Insbesondere skizziert Kapitel 4 das Design der aktiven Stufen unter Verwendung des neutralisierten Differenzpaarkonzepts sowie einer Methodik zur Optimierung der Transistor-Layoutparameter. Kapitel 5 behandelt das bekannte Problem der Leistungsverschlechterung aufgrund der Einfügedämpfung des Ausgangsanpassungsnetzwerks. Es schlägt eine ganzheitliche Figure of Merit vor, die es ermöglicht, diesen Effekt zu minimieren, und nutzt diese in einer Layout-Optimierungsmethodik für transformatorbasierte Impedanzanpassungsnetzwerke. Die Analyse der Ergebnisse liefert wertvolle Erkenntnisse über den Einfluss des Metall-Stacks und des Transformatorlayouts. Das Kapitel gibt auch Richtlinien für das Design von Messstrukturen zur Charakterisierung von Transformatoren, die mithilfe von Teststrukturen im 16-nm-FinFET validiert werden. Kapitel 6 stellt das Designverfahren eines vollständigen Leistungsverstärkers dar, ausgehend von den in den vorherigen Kapiteln entwickelten Elementen. Ein 3-stufiger Prototyp bei 80 GHz und ein 2-stufiger Prototyp bei 70 GHz werden im 16-nm-FinFET und im 22-nm-FD-SOI-Verfahren hergestellt, charakterisiert und mit früheren Arbeiten verglichen. Schließlich werden in Kapitel 7 die Schlussfolgerungen der Studie präsentiert.

Abstract

State-of-the-art wireless transceivers utilize digital-intensive architectures to exploit the advantages of CMOS technologies, but they still require analog front-end components at microwave frequencies to transmit and receive signals over the air. The limited transistor speed and breakdown voltage of CMOS technologies pose significant hurdles to the design of microwave circuits, particularly of Power Amplifiers. For this reason the analog front-end has been traditionally implemented in a separate integrated circuit based on high-performance, high-cost III-V semiconductor substrates. The ever-increasing operating frequencies reaching into the millimeter-wave spectrum have lead to more severe requirements on the transistor speed but also to better feasibility of phased antenna arrays. This has made CMOS a viable alternative to III-V technologies, enabling the integration of an entire transceiver on a single chip. Although the speed of CMOS transistors has shown a positive trend with technology scaling, the voltage handling capability has been continuously degrading, rendering the design of Power Amplifiers in deeply scaled CMOS nodes increasingly challenging. Scaling down beyond the 28 nm node was made possible by the introduction of FinFET and SOI processes. The superior digital performance of FinFET and its better scaling behavior have made it the preferred choice for digital applications, whereas SOI has shown better suitability for analog circuits. This has raised the question of whether it is more advantageous to integrate the entire system on FinFET or to distribute the digital and analog functions between FinFET and SOI. This thesis aims to investigate and compare the performance of the 16 nm FinFET and the 22 nm FD-SOI processes for the design of millimeter-wave Power Amplifiers. It delves into circuit design aspects as well as device modelling topics, with special emphasis on the characterization of standalone devices such as transistors and transformers. The investigations are conducted in the E-band (60-90 GHz), a frequency range of interest for several applications, above all automotive radar. Unlike most of the existing literature on this subject, the focus of this work is on understanding the influence of the technology features on the circuit performance rather than on

proposing innovative circuit architectures for a specific technology.
The thesis is organized as follows: Chapter 1 provides an introduction on millimeter-wave wireless links along with an overview of the relevant applications. Chapter 2 discusses the fundamentals of millimeter-wave Power Amplifiers and of the enabling semiconductor technologies focusing on deeply scaled CMOS processes. Chapter 3 proposes a novel concept for the measurement of the gate resistance, which is subsequently validated with the aid of test structures in the 16 nm FinFET technology. Chapters 4 through 6 encapsulate the core of this work, which consists in the development of an algorithmic methodology for the design of millimeter-wave Power Amplifiers to enable a systematic comparison between different processes, in this case 16 nm FinFET and 22 nm FD-SOI. More specifically, Chapter 4 outlines the design of the active stages employing the neutralized differential pair concept along with a methodology to optimize the transistor layout parameters. Chapter 5 addresses the well-known issue of the performance degradation caused by the insertion loss of the output matching network. A holistic figure of merit which allows to minimize this effect is proposed and utilized in a layout optimization methodology for transformer-based matching networks. The analysis of the results provides valuable insight into the impact of the metal stack profile and of the transformer layout. The chapter also provides guidelines to design measurement structures for the characterization of standalone transformers, which are validated with the aid of test structures in 16 nm FinFET. Chapter 6 presents the design procedure of the complete Power Amplifier starting from the building blocks designed in the previous chapters. A 3-stage prototype at 80 GHz and a 2-stage prototype at 70 GHz are fabricated in the 16 nm FinFET and in the 22 nm FD-SOI processes respectively, characterized and compared to previous art. Finally in Chapter 7 the conclusions of the study are presented.

Contents

List of Abbreviations

AC	Alternate Current 62–64
ACC	Automatic Cruise Control 7
ADAS	Advanced Driver Assistance System 7
AMPS	Advanced Mobile Phone System 8
BiCMOS	Bipolar CMOS 21
BOX	Buried Oxide 21, 22
BS	Base Station 7
BSD	Blind Spot Detection 7
CAM	Channel Access Method 6, 8
CDMA	Code Division Multiple Access 6, 8
CM	Common Mode 97, 99, 100, 103, 106, 109
CMOS	Complementary Metal-Oxide-Semiconductor vii, 4, 10, 18, 20–23, 27–29, 33, 45, 59, 70, 75, 77, 108, 112, 113
CS	Common-Source 11, 14, 16, 27, 28, 30, 31, 35, 59, 99
CW	Continuous-Wave 8, 12, 14, 16, 107, 110
DC	Direct Current 10–12, 15, 37, 61, 62, 64, 72, 91, 97–100, 102, 105, 108
DIBL	Drain-Induced Barrier Lowering 21, 22
DM	Differential Mode 99–101, 103, 106, 109
DR	Data Rate 6–8
DRC	Design Rule Checking 46, 50
DUT	Device Under Test 31, 35, 42, 43, 87, 89, 91
EDGE	Enhanced Data Global System for Mobile Communication Evolution 8

EM	Electromagnetic 2, 47, 79, 86, 88, 89, 92, 97, 98, 103, 107
EMG	Electromigration 59–65
FD	Fully Depleted 21–23
FDMA	Frequency Division Multiple Access 6, 8
FEOL	Front End of Line 61
FET	Field-Effect Transistor 18, 20, 30, 86, 109, 112
FM	Frequency Modulation 8
FMCW	Frequency Modulated Continuous Wave 8, 9
FoM	Figure of Merit 16
GF	Gate First 32, 33
GL	Gate Last 32, 33
GMSK	Gaussian Minimum Shift Keying 6, 8
GSG	Ground-Signal-Ground 35, 37
GSGSG	Ground-Signal-Ground-Signal-Ground 86, 96
GSM	Global System for Mobile Communication 8
HB	Harmonic Balance 13
HBT	Heterojunction Bipolar Transistor 20
HEMT	High Electron Mobility Transistor 20
HSPA+	WHigh Speed Packet Access + 8
IC	Integrated Circuit 1, 4, 5, 18, 20, 21, 58, 95, 99
IL	Insertion Loss 67, 70, 75, 78, 79, 81, 83, 84
ITR	Impedance Transformation Ratio 27, 28, 48, 67, 70–72, 74, 79
ITRS	International Technology Roadmap for Semiconductors 16, 48, 108, 112
LC	Inductive-Capacitive 61, 67, 70–72, 74–79, 81
LNA	Low-Noise Amplifier 18, 31
LOS	Line of Sight 2, 3, 7
LRR	Long-Range Radar 9
LS	Large Signal 99
LTE	Long-Term Evolution 8

MAG	Maximum Available Gain 47, 48
MIM	Metal-Insulator-Metal 100
MIMO	Multiple-Input Multiple-Output 6, 7
mmW	Millimeter Wave 1–12, 15, 17–24, 27–29, 31, 38, 43, 45, 67, 70, 72, 75, 86, 95, 96, 99, 108
MNW	Matching Network 13, 27–29, 45, 48, 49, 53, 57, 58, 61, 67–73, 75–86, 92, 93, 95–97, 106, 109
MOM	Metal-Oxide-Metal 46, 100
MOS	Metal-Oxide-Semiconductor 11, 32, 34, 43
NDP	Neutralized Differential Pair 46, 47, 49, 52, 54, 58, 61–66, 96, 99, 103
NF	Noise Figure 45
nMOS	n-channel Metal-Oxide-Semiconductor 21
NR	New Radio 6, 8
OFDMA	Orthogonal Frequency Division Multiple Access 6, 8
PA	Power Amplifier 10–20, 22, 23, 27–29, 31, 45–51, 53–59, 61, 65, 67, 68, 70, 72, 75–77, 79–86, 92, 93, 95–110, 112, 113
PAE	Power-Added Efficiency 11, 14, 16, 17, 104
PCell	Parametric Cell 23, 24, 50, 52, 57, 79, 80, 86, 107
PD	Partially Depleted 21, 22
PDK	Process Design Kit 23, 24, 36, 46, 97, 113
PLL	Phased-Locked Loop 18
PMCW	Phase Modulated Continuous Wave 8
pMOS	p-channel Metal-Oxide-Semiconductor 21
QAM	Quadrature Amplitude Modulation 6, 8
QPSK	Quadrature Phase Shift Keying 6, 8
RC	Resistive-Capacitive 33, 60, 103, 107

RF	Radio Frequency 1, 11, 15, 18–20, 22–26, 30, 33, 35–39, 43, 45, 46, 52, 61, 64, 67, 86, 93, 94, 96–99, 103, 104, 106, 108, 113
RFFE	Radio Frequency Front-End 3, 4, 18
RMG	Replacement Metal Gate 32
RMS	Root-Mean-Square 60
RX	Receiver 2–4, 6
SH	Self Heating 60, 61
SNR	Signal-to-Noise Ratio 6
SOA	Safe Operating Area 18, 58, 59
SOI	Silicon-on-Insulator 21–23, 28, 59, 109
SOLT	Short-Open-Load-Thru 38
SP	S-parameters 100
SRF	Self-Resonant Frequency 23, 67, 72, 73, 81–83, 86, 91, 93, 96
SRR	Short-Range Radar 9
TDMA	Time Division Multiple Access 6, 8
TL	Transmission Line 72
TRX	Transceiver 3, 5, 18
TX	Transmitter 2, 4, 6, 96
UMTS	Universal Mobile Telecommunications System 8
VCO	Voltage-Controlled Oscillator 18
VNA	Vector Network Analyzer 37, 38, 81, 91, 96, 102, 104
WCDMA	Wideband Code Division Multiple Access 8
WF	Work Function 32, 33
WLAN	Wireless Local Area Network 5
WPAN	Wireless Personal Area Network 5

1 Introduction

1.1 The millimeter wave range

The terms Radio Frequency (RF), Microwave and Millimeter Wave (mmW) refer to different subsets of the radio waves, which cover the frequency range from a few Hz up to 300 GHz. Although these terms are widely used in the literature on Integrated Circuit (IC) design, the boundaries of the corresponding ranges are often not well-defined or vary across different authors. According to [GP17], the RF frequency range is the portion of the spectrum between a few MHz and 1 GHz, the microwave range between 1 GHz and 30 GHz and the mmW range between 30 GHz and 300 GHz, as summarized in Figure 1.1.

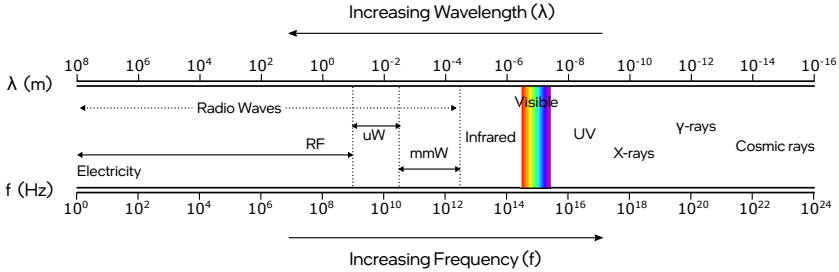


Figure 1.1: Electromagnetic spectrum.

The attractiveness of mmW frequencies lies in the availability of large, relatively unutilized portions of spectrum, which make wide bandwidths feasible. This in turn results in higher data rates, as shown by Shannon's theorem [Sha49]. This states that the channel capacity C, that is the maximum theoretically achievable data rate, is given by:

$$C = BW \times \log_2 (1 + SNR) \tag{1.1}$$

where BW is the bandwidth of the channel and SNR is the signal-to-noise ratio of the signal. Since C depends linearly on BW and logarithmically on SNR, it is easily concluded that increasing BW is the most convenient way to increase the data rate.
The main disadvantage of mmW propagation is that the geometrical attenuation of the Electromagnetic (EM) waves is proportional to f^2, where f is the frequency. Moreover, while below 10 GHz the air medium is approximately attenuation-free for EM waves, above 10 GHz the atmospheric attenuation $\alpha_{atm}(f)$ becomes considerable and frequency-dependent [MP05]. The total path loss L_{path} can be expressed as [SPD+16]:

$$L_{path} = 10\log_{10}\left(\frac{4\pi Df}{c_0}\right)^2 + \alpha_{atm}(f)D \tag{1.2}$$

where D is the distance between the Transmitter (TX) and the Receiver (RX) and c_0 is the speed of light in a vacuum. As displayed in Figure 1.2, α_{atm} shows several peaks over frequency, which are caused by various resonance effects. This restricts the usage of mmW frequencies to short-range applications and limits the usable bands to the intervals between the peaks to avoid excessive attenuation.

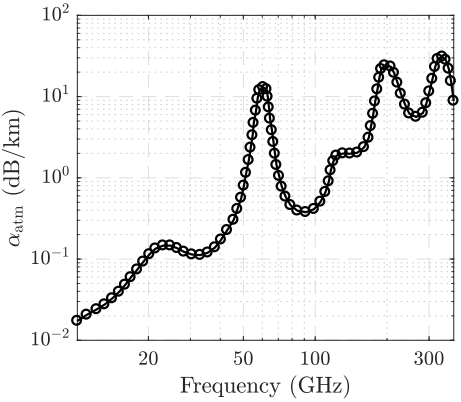


Figure 1.2: Atmospheric attenuation of EM waves at sea level.

This theory assumes Line of Sight (LOS) propagation, that is free-space prop-

agation between two points. However in virtually any real scenarios the propagation is of non-LOS type due to the presence of obstacles along the path, which results in multipath propagation effects and additional losses.
The mmW spectrum can be divided into a number of partially overlapping frequency bands, each identified by a letter, as shown in Table 1.1 [MP05].

Band	Frequency Range (GHz)
K_u	12-18
K	18-27
K_a	26.5-40
Q	33-50
U	40-60
V	50-75
E	60-90
W	75-110
F	90-170
D	110-170
G	140-220

Table 1.1: List of mmW frequency bands.

1.2 Phased Antenna Arrays

The most widespread method to counter the high path loss at mmW frequencies it is to employ a phased-array antenna concept [SCS+18], in which the Radio Frequency Front-End (RFFE) circuitry of the Transceiver (TRX) and the corresponding antennas are replicated N times, as shown in Figure 1.3a. The main advantage of this approach is that the direction of the beam can be electronically steered towards a desired target by adjusting the phase relation among the various antenna elements. This operation, which goes under the name of "beamforming", can be implemented in the analog or digital domain or by means of a hybrid approach [AKS+18]. In Figure 1.3a analog beamforming is considered, which includes a phase shifter after each RFFE block.
Applying Friis transmission equation [Poz11] to a system consisting of the phased-array TRX under analysis (Figure 1.3a) in transmission mode and another generic TRX in receiving mode, one finds that the minimum output power $P_{TX,min}$ in dBm from each RFFE block which can be detected by the RX satisfies the following link-budget equation [SPD+16]:

$$P_{tx,min}|_{dBm} = 10\log_{10}(k_B T \times 10^3 \times BW_{rx}) + NF_{rx}(f_0) + SNR_{min} + L_{fe,rx}(f_0) + \\ - G_{rx}(f_0) + L_{path}(f_0) + L_{pol} + L_{fe,tx}(f_0) - G_{tx}(f_0) \tag{1.3}$$

where k_B is Boltzmann's constant, T the absolute temperature in Kelvin, BW_{rx} the bandwidth of the RX, NF_{rx} the noise figure of the RX, SNR_{min} the minimum allowed signal-to-noise ratio for the received signal and L_{path} is defined in (1.2). Furthermore L_{pol} is the loss caused by the polarization mismatch between the TX and RX antennas, $L_{fe,tx}$ and $L_{fe,rx}$ are the losses caused by the front-end components of the TX and RX chains respectively. Finally G_{tx} and G_{rx} are the gain of the transmitting and receiving antennas respectively.
Using a phased array as transmitting antenna, one obtains:

$$P_{tx,min}|_{dBm} = 10\log_{10}(k_B T \times 10^3 \times BW_{rx}) + NF_{rx}(f_0) + SNR_{min} + L_{fe,rx}(f_0) + \\ - G_{rx}(f_0) + L_{path}(f_0) + L_{pol} + L_{fe,tx}(f_0) - G_{a,tx}(f_0) - 10\log_{10}(N^2) \tag{1.4}$$

where $G_{a,tx}$ is the gain of a single antenna element and the term N^2 results from the fact that the overall antenna gain and transmitted power are N times larger than those of a single antenna element. The conclusion that the transmitted power increases by a factor N^2 is however not correct, since one should compare to a single antenna with the same physical size of the array, which would result in the same gain and power. As a matter of fact the key advantage of the array is that the reduced output power requirements on each TX unit block allow in many applications to use deeply scaled CMOS technologies, as explained in section 2.2. This mitigates the concerns about the increased area occupation and power consumption of the RFFE. Furthermore, the phased array concept allows to electronically steer the direction of the beam. It is noteworthy that phased arrays become feasible at mmW frequencies because the required size and spacing of the antenna elements are compatible with the size of the IC. In state-of-the-art systems the antenna array is typically implemented by means of patch elements built on the package of the IC [ZGZ+21], as shown in Figure 1.3b, which minimizes the footprint and the interconnection losses.

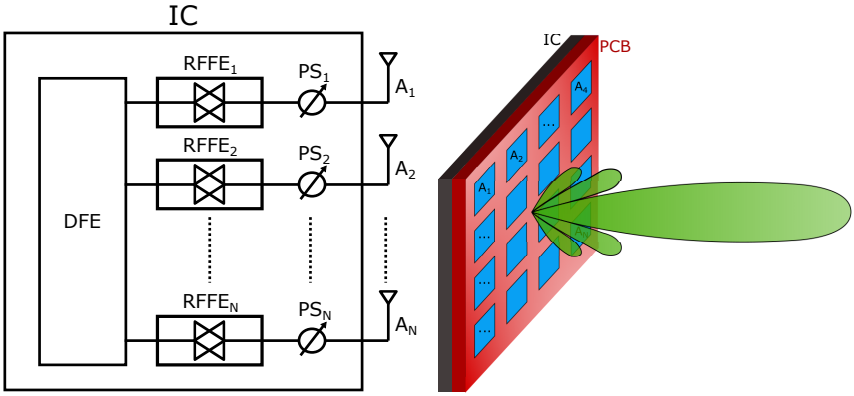


Figure 1.3: Illustration of a TRX concept based on phased antenna arrays, including (a) the circuit schematic of the TRX IC and (b) a view of a patch antenna array implemented on the package of the IC.

1.3 Applications at millimeter-wave frequencies

In the last 2-3 decades the interest of analog circuit designers has progressively shifted towards the mmW frequency range due to the ermergence of countless new applications. The main fields of interest are Wireless Local Area Network (WLAN) and Wireless Personal Area Network (WPAN) [RKS17], mobile backhaul [DOA+17], automotive radar [Sch05], satellite communications [KLM+20] and mmW imaging [WCC19]. Cellular communications have also been affected through the introduction of the mmW bands in the 5G standard. In the rest of this chapter cellular communications and automotive radar are discussed at some length due to their relevance for this research work.

1.3.1 Cellular Communications

Cellular communications is by far the most popular field due to its impact on the everyday life of virtually anyone. The last three decades have witnessed an unprecedented development of the communication standards, driven by the need to transfer increasing volumes of data with the lowest possible latency. A rapid

evolution has come about from the earliest purely analog first generation (1G) in the 80s with maximum achievable Data Rate (DR) of $DR_{max} = 9.6\,kb/s$, all the way to the latest 5G New Radio (NR) in the late 2010s with $DR_{max} = 20\,Gb/s$ [BB+10,MSM15]. This impressive evolution has been attained by continuously improving the following aspects:

- The Channel Access Method (CAM), namely the utilization of the air propagation medium. Simultaneous communication channels can be enabled transmitting information at different frequencies using Frequency Division Multiple Access (FDMA) or at different times by means of Time Division Multiple Access (TDMA). Another technique is the Code Division Multiple Access (CDMA), along with its 3G evolution CDMA2000, which is based on the orthogonal coding of the information. Starting from 4G the dominant standard has become Orthogonal Frequency Division Multiple Access (OFDMA), which combines aspects of both TDMA and FDMA.
- The Modulation Scheme, namely the way the information is encoded on the carrier signal, which starting from 2G is based exclusively on digital techniques. Depending on the utilized modulation scheme, a different number of symbols can be transmitted or received in a given bandwidth. Gaussian Minimum Shift Keying (GMSK), used only in 2G, can transmit 2 symbols, whereas Quadrature Phase Shift Keying (QPSK), which has become widely used starting with 2G, can transmit 4 symbols. Finally Quadrature Amplitude Modulation (QAM), the most widespread scheme in the 4G and 5G standards, can transmit 16, 64, 128, 256 or even 1024 symbols. A limit to the number of symbols is posed by the Signal-to-Noise Ratio (SNR) of the received signal, which should be large enough to allow for correct demodulation.
- Multiple-Input Multiple-Output (MIMO) [HGPR+16], that is the utilization of multiple antennas at the TX and RX to implement beamforming and spatial multiplexing [CFH+10]. The term "massive MIMO" is often used in conjunction with mmW communications to indicate the presence of large antenna arrays, which enable hybrid approaches between beamforming and spatial multiplexing (see Figure 1.4).
- Polarization diversity, that is the transmission of two different signals over a single antenna exploiting the two orthogonal polarizations.

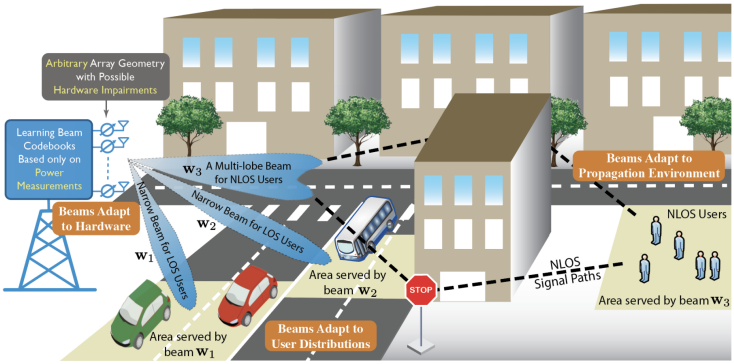


Figure 1.4: 5G Base Station (BS) utilizing massive MIMO to serve multiple LOS and non-LOS users in a scattering-rich environment (reprinted from [ZAA21] ©2022 IEEE).

While the generations from 1G to 4G operate mostly in the sub-6 GHz range, the latest 5G, currently under deployment, makes use for the first time of frequency bands in the mmW range. Interestingly, the most important mmW bands of the 5G standard are located mostly in the K_a band, between 28 GHz and 42 GHz, that is between the first and second attenuation peak of Figure 1.2. Significant research efforts are currently in progress to enable 6G communications, which are expected to utilize operating frequencies in the sub-Terahertz band and reach DR beyond 100 Gb/s [YXXL19].

1.3.2 Automotive Radar

A radar is a system whose goal is to identify the position and velocity of an object. In the automotive context it is used mostly for the Automatic Cruise Control (ACC), Advanced Driver Assistance System (ADAS) and Blind Spot Detection (BSD) functions and is playing a pivotal role in the advance of autonomous driving. In many cases the automotive radar is monostatic, that is the same antenna performs the transmitting and receiving functions, as shown in Figure 1.5a. Applying Friis transmission equation one can determine the maximum distance R_{max} at which an object with radar cross section σ can be detected:

Generation	Standards	CAM	Frequency (GHz)	DR_{max} (Mbps)	Modulation Schemes
1G	AMPS	FDMA	0.85	0.0096	FM
2G	GSM, EDGE	TDMA, CDMA	0.38-1.9	0.056	GMSK, QPSK
3G	UMTS (WCDMA, HSPA+), CDMA2000	CDMA	0.7-2.6	14.7	QPSK
4G	LTE	OFDMA	0.7-5.9	150	QPSK, 16/64 QAM
5G	NR	OFDMA	0.7-71	20000	QPSK, 16/64/256 QAM

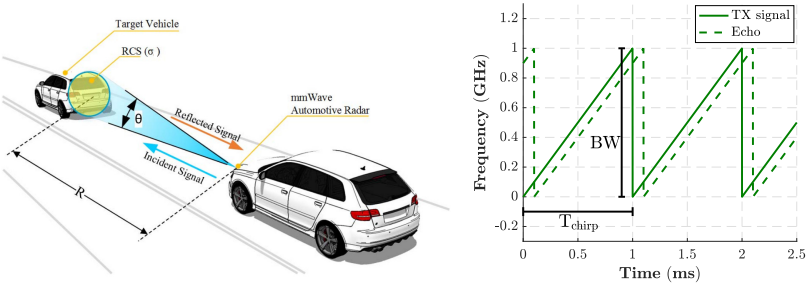
Table 1.2: Summary of the 5 generations of mobile communication standards

$$R_{max} = \sqrt[4]{\frac{P_{tx}\sigma G^2 c_0^2}{P_{rx,m}(4\pi)^3 f^2}} \tag{1.5}$$

where P_{tx} is the transmitted power, $P_{rx,m}$ the minimum receiver power which can be detected and G the antenna again. The equation shows that if the aperture of the antenna is assumed constant over frequency, R_{max} is proportional to $1/\sqrt{f}$ [Sko80].

One first classification of radar systems is based on the type of signal utilized to illuminate the target: the most classical implementation uses pulsed signals, whereas more recent, fully integrated systems use the Frequency Modulated Continuous Wave (FMCW) [B+05] or Phase Modulated Continuous Wave (PMCW) [GSD+17] approaches. In the FMCW radar a target placed at distance R is illuminated using a mmW Continuous-Wave (CW) carrier with linear frequency modulation of period T_{chirp} and excursion $\Delta f = BW$, as shown in Figure 1.5b.

The distance R of the target can be computed from the frequency f_{beat} of the beat between the transmitted signal and the reflected echo:



(a) (b)

Figure 1.5: Automotive radar: (a) illustration of operating principle (reprinted from [AHEK+17] ©2018 IEEE) and (b) typical waveforms employed in FMCW radars.

$$\mathrm{f_{beat}} = \frac{\mathrm{BW}}{\mathrm{T_{chirp}}} \frac{2R}{c_0} \tag{1.6}$$

Since the maximum beat period which can be resolved is $\mathrm{T_{beat}} = \mathrm{T_{chirp}}$, the minimum measurable beat frequency is $\delta \mathrm{f_{beat}} = 1/\mathrm{T_{chirp}}$ and the range resolution δR of the radar is given by:

$$\delta \mathrm{R} = \frac{\mathrm{c_0}}{2\mathrm{BW}} \tag{1.7}$$

This equation shows that the resolution improves as the BW increases, which explains why also the operating frequencies of the radar have been progressively pushed towards the mmW range. Based on the maximum detectable distance, automotive radar systems can be classified into Short-Range Radar (SRR) and Long-Range Radar (LRR). The former typically operates at 24 GHz, the latter at 77 GHz, which fall into the K-band and E-band respectively. Recent research developments show increasing interest in the D-band for next-generation automotive radars [AKE+20].

2 Basics of Power Amplifiers and Enabling Technologies

The emerging applications mentioned in Chapter 1 have brought new and significant challenges to the implementation of microwave circuits and systems. One key aspect is the selection of the proper semiconductor technology, which has to take into account performance as well as cost considerations. While for digital circuits CMOS technologies are an obvious choice, mmW front-ends can be implemented in a variety of different technologies ranging from III-V compound semiconductors to CMOS in its multiple variants to silicon bipolar. From the performance perspective one could claim that the technology choice is mainly driven by the Power Amplifier (PA). This is indeed the most challenging block in the transmitter chain of a transceiver, because it has to generate and sustain high power levels with good efficiency and often wide bandwidth, while meeting potentially stringent linearity requirements [WS15]. While technology selection is crucial, the employment of suitable circuit design techniques is at least equally important for a successful design. This Chapter outlines the basic concepts of solid-state PA design at mmW frequencies and presents a short review of the main technology options along with their strengths and weaknesses. Special attention is devoted to deeply scaled CMOS technologies, which constitute the main focus of this work, and to the design techniques which allow to overcome their limitations.

2.1 Basics of Power Amplifiers

2.1.1 Terminology and Definitions

A PA is an active circuit designed to amplify the power of a signal with bandwidth BW from a level P_{in} to a level P_{out}, with $P_{out} > P_{in}$. To do so, a Direct

Current (DC) power P_{DC} is absorbed from a supply, which provides the active devices in the circuit with the necessary bias current. For CMOS technologies, which are the focus of this work, the supply voltage is denoted by V_{DD} and the bias current by I_d, where the subscript "d" refers to the drain terminal of a Metal-Oxide-Semiconductor (MOS) transistor. The most important figures of merit of a PA are the output power P_{out}, the gain $G = P_{out}/P_{in}$, the drain efficiency $\eta = P_{out}/P_{DC}$ and the linearity. At mmW frequencies the Power-Added Efficiency (PAE) is typically used instead of η to take into account the effect of the limited G of the amplifying stages. It is defined as:

$$\mathrm{PAE} = \frac{P_{out} - P_{in}}{P_{DC}} = \frac{P_{out}}{P_{DC}}\left(1 - \frac{1}{G}\right) = \eta\left(1 - \frac{1}{G}\right) \tag{2.1}$$

2.1.2 Loadline and Loadpull Analysis

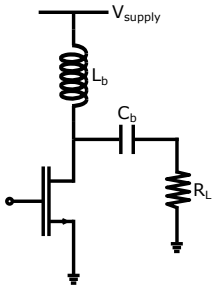


Figure 2.1: Simple Common-Source (CS) stage considered in the loadline theory.

A PA normally operates with large signal levels, which bring the active devices to some extent out of the linear operating region. The standard circuit characterization by means of the S-parameters is therefore insufficient and large-signal simulation and characterization techniques are required. Considering the simple CS stage of Figure 2.1, where L_b and C_b form a bias-T circuit to decouple the DC bias from the RF signal, the output power is given by [RS06]:

$$P_{out} = \frac{V_{supply}^2}{2R_{L,opt}} \tag{2.2}$$

where V_{supply} is the supply voltage and $R_{L,opt}$ is the optimum load resistance which maximizes the output voltage and current swing. This result is based on the highly simplistic loadline theory [Cri06], which combines the I-V characteristics of the transistor with that of the purely resistive load R_L, as shown in Figure 2.2a. The most important assumption is that the knee voltage V_{knee} of the DC characteristics is equal to 0 V, as shown in Figure 2.2b, so that the transistor remains in the saturation region over the entire swing of the output voltage. Since the DC model of the transistor does not include any reactive components, the optimum load can be assumed to be a pure resistance. This is normally a reasonable hypothesis at frequencies of a few GHz, but certainly not in the mmW range. Under all these assumptions, if the output voltage signal $|v_{out}|$ swings between 0 V and $2V_{supply}$ and $R_L = R_{L,opt}$, the output current signal is $|i_{out}| = |v_{out}|/R_{L,opt}$, from which (2.2) follows immediately.

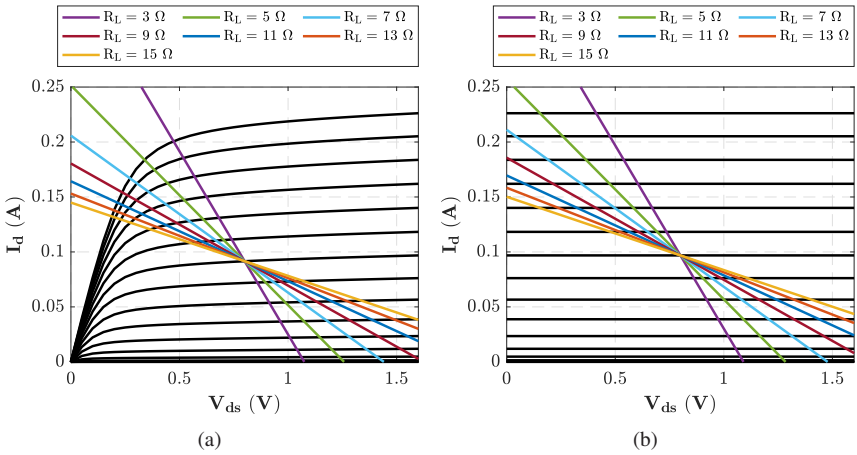


Figure 2.2: Typical loadline plots of a PA with (a) full and (b) approximated (V_{knee} = 0 V) DC characteristics of the active device.

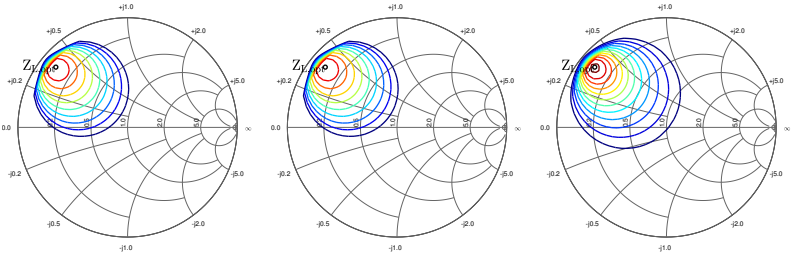
In a pratical design scenario the large-signal behavior is studied using the full transistor model, which includes the dynamic and non-linear effects [SSKJ87]. The analysis conducted in the following is based on such device models and holds for a CW excitation at a fixed operating frequency in the mmW range. Due to the terminal parasitic capacitances, the active device presents significant capacitive reactance at the output, namely:

$$Z_{out} = R_{out} + jX_{out} \tag{2.3}$$

with $X_{out} < 0\,\Omega$. For this reason the optimum load impedance $Z_{L,opt}$ is in general complex:

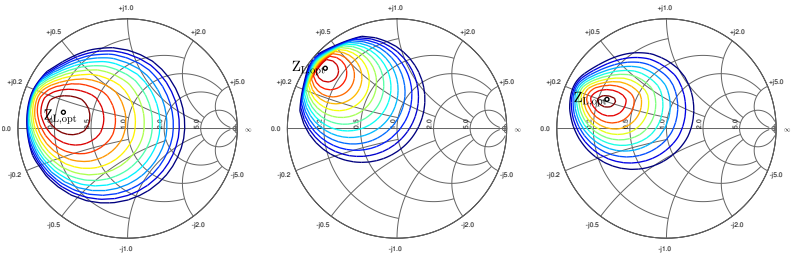
$$Z_{L,opt} = R_{L,opt} + jX_{L,opt} \tag{2.4}$$

where $X_{L,opt} = -X_{out}$ is an inductive reactance which neutralizes the output capacitance of the active device. Since R_{out} and X_{out} vary as a function of the signal level, the optimum load impedance can only be determined running a Harmonic Balance (HB) analysis with a generic load impedance $Z_L = R_L + jX_L$ and performing a two-dimensional sweep of R_L and X_L. This goes under the name of loadpull analysis and produces the loadpull circles, namely the contours of constant P_{out}, G or PAE. Depending on which of these quantities have to be optimized, the corresponding $Z_{L,opt}$ is determined [GH12]. Typical loadpull circles of P_{out}, G and PAE for a PA output stage are shown in Figure 2.3 and 2.4 for small-signal and large-signal conditions respectively. Figure 2.3 shows that as long as small-signal conditions are maintained, the loadpull circles have cicular shape and the optimum impedances for P_{out}, G and PAE coincide. Figure 2.4 instead shows that as the drive level increases and brings the PA out of the linear operating region, the curves tend to become elliptical and the optimum impedance values progressively depart from each other [Cri06]. In this scenario all the quantities are evaluated at a fixed compression point to ensure that the PA is always in the same "amount" of large-signal drive in the various loading conditions which are compared. In this work the 3-dB compression point is chosen, as indicated by the "3dB" subscript in Figure 2.4. The output Matching Network (MNW) is normally designed in such a way to provide the PA with the optimum impedance level determined by the loadpull analysis, as discussed at length in Chapter 5.
Once fixed the load impedance, the behavior of P_{out}, G and PAE as a function of P_{in} can be simulated. One typical outcome is shown in Figure 2.5: for low P_{in}, G is constant and equal to the small-signal value G_{ss} but at some point it starts decreasing or "compressing", where the compression level is defined as $C(P_{in}) = G_{ss} - G(P_{in})$. This happens because the output voltage or current of the active devices fall out of the saturation region and start clipping. As a consequence the PA approaches the maximum P_{out} it can generate with the



(a) P_{out}. (b) G. (c) PAE.

Figure 2.3: Typical loadpull curves for a PA output stage in small-signal conditions.



(a) $P_{out,3dB}$. (b) G_{3dB}. (c) PAE_{3dB}.

Figure 2.4: Typical loadpull curves for a PA output stage at the 3-dB compression point.

given V_{supply}, the so-called saturation power P_{sat}. As a result of these different behaviors, the PAE normally shows a peak (PAE_{peak}) when the compression level is equal to a few dB, as shown in Figure 2.5.

2.1.3 Operating Class

A linear or quasi-linear PA excited by a CW stimulus is said to operate in class A, AB or B depending on the fraction of the period during which the active devices conduct [Cri06]. In class-A the transistors are turned on during the entire signal period, in class-B only during half of it. All the intermediate conditions fall into the so-called class-AB. The time-domain waveforms of the gate-to-source voltage $V_{gs}(t)$ and drain current $I_d(t)$ for a single-stage CS PA operating at $f_0 = 1\,\text{GHz}$ are shown in Figure 2.6 for the three scenarios.

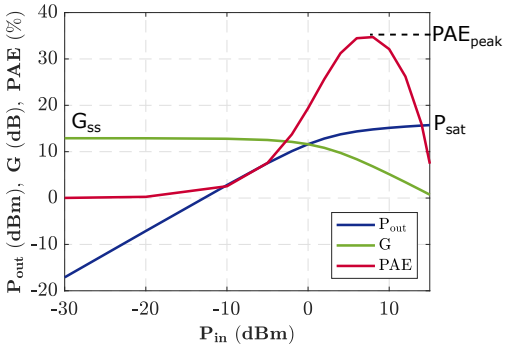


Figure 2.5: Typical large-signal behavior of a mmW PA.

The operating frequency is purposedly chosen in the low RF range because it allows to observe the clipping behavior of $I_d(t)$. Due to the limited f_t and f_{max}, at mmW frequencies the transistor can not generate enough harmonics to reproduce exactly the sharp transition from above threshold to below threshold. This explains also why switching PA modes such as classes D, E and F are normally not used at mmW frequencies [CK14].
The operating class is dictated by the gate-to-source bias voltage $V_{gs,0}$ of the transistor relative to its threshold voltage V_t. For $V_{gs,0} = V_t$ the amplifier operates in class B, whereas for $V_{gs,0} > V_t$ it operates either in class AB or in class A depending on the value of P_{in}. It operates namely in class A if the waveform $V_{gs}(t)$ never drops below V_t for all the applied values of P_{in}, otherwise it operates in class AB. From Figure 2.6 it is clearly visible that the utilized active device has $V_t \sim 0.2\,V$. It can be proved analytically that the maximum achievable drain efficiency η_{max} increases as one moves from class A ($\eta_{max} = 50\%$) towards class B ($\eta_{max} = 78\%$) thanks to the decreasing DC component of $I_d(t)$. As long as G is larger than approximately 10 dB, which is normally the case at RF frequencies, one has $(1 - 1/G) \geq 0.9$ and therefore PAE $\sim \eta$. However at mmW frequencies G could be as low as 10 dB or less, so that PAE can drop significantly below η. Since moving towards class-B results also in lower G as an effect of the lower f_t, at mmW frequencies it is typically more convenient to bias in class A or AB.

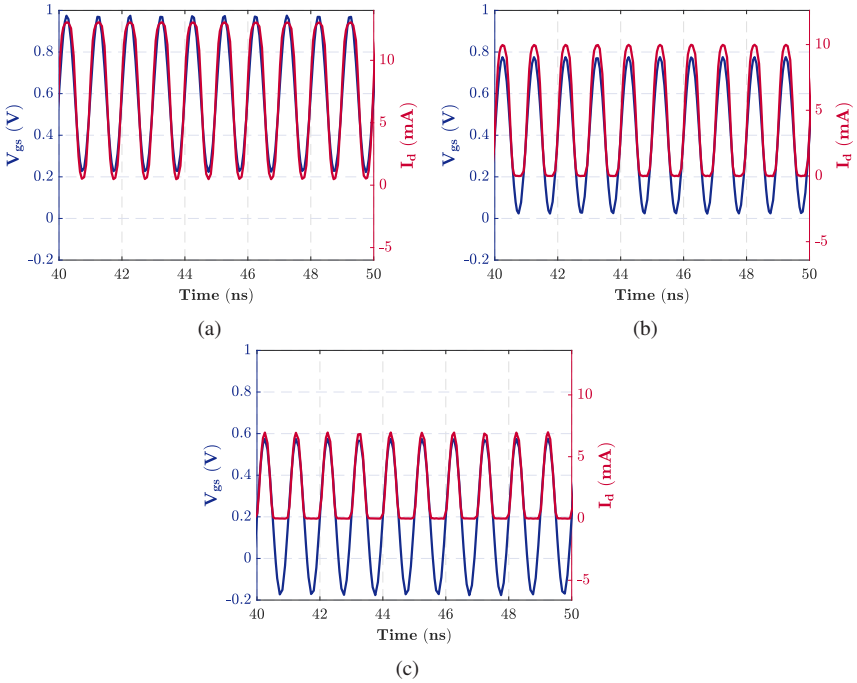


Figure 2.6: $V_{gs}(t)$ and $I_d(t)$ waveforms of a CS single-stage PA biased in (a) class A ($V_{gs,0}$ = 0.6 V), (b) class AB ($V_{gs,0}$ = 0.4 V) and (c) class B ($V_{gs,0}$ = 0.2 V), excited by a CW signal at f_0 = 1 GHz with P_{in} = −5 dBm.

2.1.4 Figure of Merit, Efficiency of Multi-Stage PA

Assessing the overall performance of a PA requires a unified way to take into account all the metrics introduced in section 2.1.1, that is P_{out}, G and PAE. As an example, designing a circuit targeting only very large P_{out} could end up in an implementation with very low G and therefore very low PAE. On the other hand one might seek a solution with very large PAE, but end up with an insufficient P_{out} for the target application. The most common criterion to assess the overall performance of a PA is the Figure of Merit (FoM) from the International Technology Roadmap for Semiconductors (ITRS) [A+05], which combines all these metrics in a single expression. It is given by:

$$FoM = G_{ss}(dB) + P_{sat}(dBm) + 10\log_{10}(PAE_{peak}(\%)) + 20\log_{10}(f_0) \quad (2.5)$$

The frequency term f_0 is critical for mmW PAs because at high frequency it becomes increasingly hard to obtain large G and PAE.
Due to the low gain of the active devices, at mmW frequencies multi-stage designs are normally required to achieve the desired amplification. Unfortunately every additional amplifying stage causes a degradation of PAE. Considering a two-stage amplifier (Figure 2.7) where each stage S_i has power gain G_i and power-added efficiency PAE_i, with $i = 1, 2$, the overall PAE (PAE_{PA}) is given by [YMYZ15]:

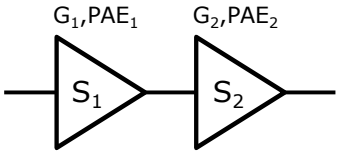


Figure 2.7: Two-stage PA chain.

$$PAE_{PA} = PAE_2 \left[1 - \frac{PAE_2 - PAE_1}{PAE_2 + PAE_1 \dfrac{G_1(G_2 - 1)}{G_1 - 1}} \right] \quad (2.6)$$

The equation shows that 1) PAE_{PA} can be at most equal to PAE_2 and 2) the larger the gain G_2 of stage S_2, the lower the impact of PAE_1 on PAE_{PA}. Adding a stage S_0 before S_1 and applying (2.6) iteratively it can be verified that the impact of each new stage on the overall PAE_{PA} becomes less and less relevant. The main conclusion is that the output stage, also called "core" throughout this work, should have large G and PAE to optimize the performance of the PA and minimize the sensitivity on the preceding stages, the so-called "drivers". For this reason the design of a PA starts always from the output stage and then proceeds backwards to the less critical drivers.

2.2 Enabling Technologies

The transistor is the fundamental building block of every IC and the reference component of any semiconductor technologies. In digital circuits it works essentially as a switch, with the output signal commuting between two logical states denoted as 0 and 1. In state-of-the art microprocessors clock frequencies of at most a few GHz are used [PPS+13]. On the other hand in analog circuits such as the RFFE of a TRX (see section 1.2) transistors work with continuous-amplitude signals, normally a sinuoidal carrier in the RF or mmW frequency range modulated by a signal with bandwidth BW which carries the information to be transmitted or received. The RFFE includes amplifiers such as PAs and Low-Noise Amplifiers (LNAs), circuits for frequency up- and down-conversion (mixers), frequency dividers and multipliers, Voltage-Controlled Oscillators (VCOs), Phased-Locked Loops (PLLs), filters and many others. The digital or analog nature of an IC is the fundamental criterion for the selection of the most appropriate semiconductor process. As mentioned in the introduction, in the digital domain the technology of choice is CMOS due to its superior performance, maturity, low cost and ease of mass production. In the analog domain instead the picture is more complicated due to the availability of several technology options with different trade-offs between cost and performance. The two broadest categories are the so-called III-V compund semiconductors, which include GaAs, GaN and InP, and the silicon-based technologies, namely CMOS and bipolar. Some of these technologies utilize Field-Effect Transistor (FET) devices, some others rely on bipolar transistors and a few of them offer even both options. In the realm of analog circuits the desired features are mostly large unity gain frequency f_t and maximum oscillation frequency f_{max} (see section 3.1). They depend, among other things, on the electron mobility μ_n in the utilized material and, as a rule of thumb, should be larger than twice the operating frequency of the circuit. Specifically for PAs, an equally important feature is the voltage-handling capability, that is the ability to sustain high output voltage swings. The reference figure of merit is the breakdown field E_{bd} or, equivalently, the off-state breakdown voltage BV_{off}, that is the smallest drain-to-source voltage for which the device is permanently damaged. In practice BV_{off} provides an upper bound for V_{supply}, which translates into a maximum P_{out} achievable by the PA (2.2). In CMOS technologies BV_{off} is not always reported because the Safe Operating Area (SOA) is more severely limited by the gate oxide breakdown, as clarified in section 4.4. Moreover $V_{supply} = V_{DD}$

is normally used, where V_{DD} is the supply voltage for digital applications. For technology selection purposes it should be kept in mind that a semiconductor material with large bandgap typically results in high BV_{off} but low μ_n [PAO+02]. Table 2.1 shows an overview of the most advanced nodes for each technology with the corresponding RF figures of merit. Additionally Figure 2.8 shows the result of a survey conducted by GeorgiaTech [WWN+20] which includes the most important mmW PAs published between the years 2000 and 2020, in which all the above-mentioned technologies are represented. These data are used throughout the rest of this section to compare the various technologies and highlight their advantages and disavantages.

Table 2.1: RF figures of merit of RF transistors in common semiconductor technologies for mmW applications.

Category	**Technology**	$\mathbf{f_t}$ **(GHz)**	$\mathbf{f_{max}}$ **(GHz)**	$\mathbf{V_{supply}}$ **(V)**	$\mathbf{BV_{off}}$ **(V)**
III-V	35 nm GaAs HBT [ARCRVR+18]	515	1000	-	2
III-V	20 nm GaN HEMT [SRT+13]	450	600	-	10
III-V	130 nm InP DHBT [UPR+11]	520	1100	-	3.5
Bipolar	90 nm SiGe HBT [PAG+14]	505	720	-	1.6
CMOS	45 nm planar [LJW+07]	350	280	1.2	-
CMOS	22 nm FinFET [LCR+20]	300	450	1	-
CMOS	22 nm FD-SOI [OLC+18]	350	370	0.8	-

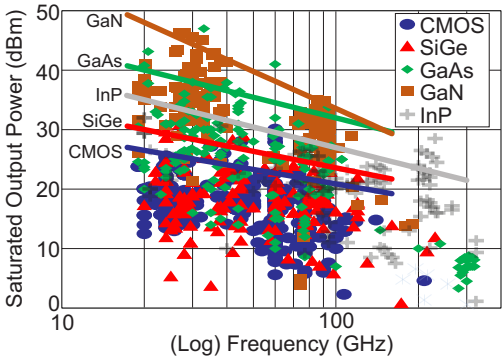


Figure 2.8: Survey of P_{sat} of mmW PAs published in the time frame 2000-2020 [WWN+20] (reprinted from [CQP+20] ©2020 IEEE).

2.2.1 III-V Compund Semiconductors

III-V compound semiconductors have been for many years the technology of choice for RF and mmW applications thanks to their large μ_n and BV_{off}. Among the processes listed in Table 2.1, GaAs and InP utilize mostly the Heterojunction Bipolar Transistor (HBT), whereas GaN is based on the High Electron Mobility Transistor (HEMT), a special type of FET, but in principle both types of structures are feasible in each technology. III-V semiconductors are still widely used for PAs which require very high output power and do not benefit from the usage of phased arrays (see section 1.2), for instance in sub-6GHz cellular applications. Figure 2.8 shows that III-V compounds are in this regard much better than silicon-based technologies. More specifically, GaN shows the highest output power but tends to converge with GaAs as the operating frequency increases above 100 GHz. Consistently with this observation, Table 2.1 shows that GaN has by far the largest BV_{off} but also the worst f_t/f_{max} among III-V technologies [SRT+13]. InP is not as good as GaN in terms of voltage handling capabilities but is the absolute best in terms of f_t/f_{max}, with reported values of 520 GHz/1100 GHz [UPR+11]. Although GaAs has achieved lately almost comparable performance [ARCRVR+18], most PAs in the sub-THz range up to 300 GHz published to this day are designed in InP. The main disadvantages of III-V processes are the high cost and the non suitability for mass production. Moreover the presence of separate ICs for the digital and analog functions leads invariably to considerable losses at the interface between the two.

2.2.2 Silicon-based Technologies: CMOS and Bipolar

CMOS technologies have been historically conceived and developed exclusively for digital applications, since the limited μ_n of Si made them unsuitable for RF and mmW applications. The continuous scaling to smaller gate length L_g, however, allowed to overcome this limit, as shown by the equation below:

$$f_t = \frac{\mu_n E}{2\pi L_g} = \frac{v_{n,sat}}{2\pi L_g} = \frac{1}{2\pi \tau_t} \tag{2.7}$$

where E is the electric field across the channel, $v_{n,sat}$ the electron saturation velocity and τ_t the channel transit time of the electrons. Starting from the

0.18 µm node in the 2000s it became clear that f_t and f_{max} values compatible with mmW operation could be attained [SHBR05]. This triggered a huge research effort to optimize the technology for these applications, with the intent of fabricating an entire transceiver on a single IC. Unfortunately the scaling comes also with a strong decrease of BV_{off} [RZN19, TF03], which poses severe limitations to the voltage handling capability. Bipolar silicon technologies such as SiGe are often used instead of CMOS as they allow to attain larger P_{out} thanks to the superior BV_{off} of bipolar devices. Most SiGe processes offer the Bipolar CMOS (BiCMOS) option, that is the integration of CMOS devices on the same substrate, to combine the advantages of the two technologies. Despite the cost being lower than that of III-V semiconductors, it is still significantly larger than that of CMOS and mass production is not equally easy. Nonetheless SiGe BiCMOS is nowadays one of the most popular mmW technologies for applications whose output power specifications cannot be met by conventional CMOS.

2.2.3 Deeply Scaled CMOS Technologies

One of the key concept of CMOS technologies for digital applications is the presence of two complementary transistor types, the n-channel MOS (nMOS) and the p-channel MOS (pMOS). For mmW circuits nMOS devices are mostly of interest, whereas pMOS have been traditionally avoided due to the lower mobility of the holes compared to the electrons ($\mu_h < \mu_n$). Recently the performance of pMOS devices was significantly improved utilizing strain engineering, so that their usage for these kind of applications has become possible [JBA16]. The standard CMOS variant is the so-called planar bulk CMOS, shown in Figure 2.9a. With the continuous downscaling this concept started to approach its limits due to the exacerbated Drain-Induced Barrier Lowering (DIBL) and increased subthreshold current I_{sub}, which pushed the off-state power consumption of digital circuits to unacceptable levels. In order to continue the scaling process beyond the 28 nm node [Kuh12], two new approaches were devised to mitigate those effects, namely Silicon-on-Insulator (SOI) [KN14] and FinFET [JCV+09]. SOI devices in their Partially Depleted (PD) and Fully Depleted (FD) variants, are obtained by adding a Buried Oxide (BOX) layer to isolate the channel region from the substrate and block the current leakage through the substrate itself (Figure 2.9b). In FinFET devices instead the planar channel

region is replaced by a three-dimensional structure made up of a number of discrete fins, which are wrapped by a triple gate electrode for better electrostatic control of the channel (Figure 2.9c). It should be noticed that the SOI and FinFET structures are not mutually exclusive, since SOI affects only the substrate region and FinFET only the channel region. The most correct naming for the mentioned CMOS processes is therefore bulk planar, SOI planar and bulk FinFET. More recently, FinFET SOI technologies have been proposed [ZCT+16] to combine the advantages of the two approaches, but to this day they have not yet been widely adopted.

FinFET processes are particularly suitable for digital applications due to the superior electrostatic control of the channel, analog gain, DIBL, subthreshold slope, device density and scaling behavior. However for mmW applications they are disadvantaged by their three-dimensional structure. The increased parasitic capacitances [TRML+12] and gate resistance lead indeed to degraded f_t and f_{max}, so that further device optimization is required. Furthermore the larger current density in the metal interconnects results in higher sensitivity to Electromigration, as explained extensively in section 4.4.1. On the other hand, in SOI processes the BOX reduces the parasitic capacitances towards the bulk, resulting in relatively large f_t/f_{max}. Moreover it eliminates the parasitic diode between the source/drain (S/D) implant regions and the bulk, which eases the implementation of stacked PA architectures (see section 2.3). Thanks to the particularly thin BOX, in FD-SOI technologies a so-called back-bias voltage V_{bb} can be applied between the well and the substrate, which allows for a wide-range tuning of V_t [ZLO+21]. With $V_{bb} > 0\,V$ a larger V_t is obtained, whereas $V_{bb} < 0\,V$ results in lower V_t.

One major weakness of CMOS technologies at RF and mmW frequencies is that they require low substrate resistivity ρ_{sub} [NRS+22, BYC+03], typically a few $\Omega\,cm$, to mitigate latch-up effects [Hu84]. The main consequence is that the quality factor Q of the passives is degraded due to the strong electrical coupling to the substrate. A large ρ_{sub} is feasible only in PD-SOI processes [RNN+22], with reported values in the order of a few $k\Omega\,cm$.

In order to enable interconnections with low resistivity and passives with satisfactory Q, CMOS technologies optimized for RF applications typically offer metal stack options with one or more ultra-thick metal layers. These are normally located at the top of the stack to minimize the electrical coupling to the substrate. Another important feature is the vertical separation between the metal layers. It plays a major role for vertically-coupled transformers (see

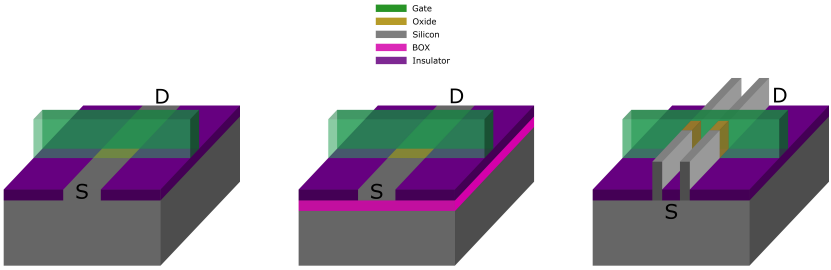


Figure 2.9: Illustration of the transistor structure in the main CMOS flavors: (a) bulk planar, (b) SOI planar and (c) bulk FinFET.

section 5.2.2), as it influences the magnetic coupling between primary and secondary and the Self-Resonant Frequency (SRF).

2.2.4 The 16 nm FinFET and 22 nm FD-SOI processes

In this thesis the most utilized technology is a 16 nm FinFET process [WLC+13], abbreviated as "16FF" in the text. It is used for the gate resistance and transformer characterization test structures in Chapter 3 and 5 respectively as well as for the three-stage PA prototype described in Chapters 4, 5 and 6. The second utilized technology is a 22 nm FD-SOI process [ZLO+21], abbreviated as "22SOI", which is used for the two-stage PA prototype described also in Chapters 4, 5 and 6. From the point of view of mmW circuit design, the most critical features of a technology are the availability of accurate high-frequency active and passive device models in the Process Design Kit (PDK) and of a suitable metal stack. In the rest of this section this information is provided for the 16FF and 22SOI processes.

The 16FF process offers a layout Parametric Cell (PCell) for the RF transistor which allows to vary the number of fins N_{fins}, number of fingers N_{fing}, gate length L_g and gate pitch PP. The multiplicity M, that is the number of transistor units connected in parallel, is also important for the design but is not a parameter of the PCell. A graphical representation of the geometrical features of the RF transistor is provided in Figure 2.10a. The parameter PP can take on the values

$1 \times PP_{ref}$, $1.07 \times PP_{ref}$ and $1.44 \times PP_{ref}$, where PP_{ref} is the smallest available value and $1.44 \times PP_{ref}$ was introduced by the foundry as enhanced option for mmW applications. The gate length $L_g = N \times L_{g,ref}$ can take on several values, where $L_{g,ref}$ is the smallest possible value and N can be equal to 1, 1.125, 1.25 and 1.5. Values of N larger than 1.5 are also available but not of interest for the application at hand because they cause a degradation of the channel transit time τ_t (2.7).

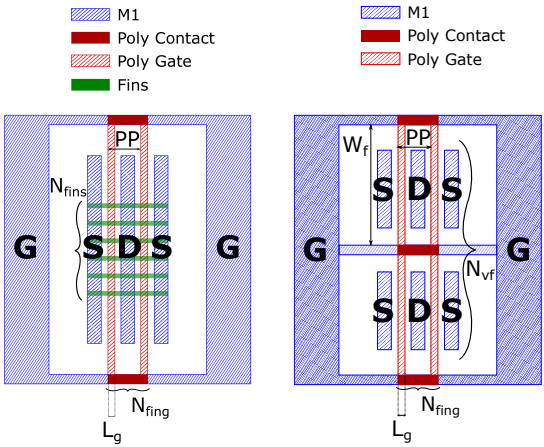


Figure 2.10: Parameters of the RF transistor layout PCell in the (a) 16FF and (b) 22SOI processes.

The 22SOI process has been developed and advertised as mmW technology, and indeed the PDK offers a mmW transistor PCell with several convenient features. The main parameters are the finger width W_f, number of vertical fingers N_{vf}, number of fingers N_{fing}, gate length L_g and gate pitch PP (see Figure 2.10b). The gate pitch PP can take on the values $P = 1 \times PP_{ref}$, $2 \times PP_{ref}$ and $3 \times PP_{ref}$, where PP_{ref} is the smallest available value but not the same as in 16FF. The transistor PCell has the option to split the gate fingers longitudinally in a number N_{vf} of sections (NREP in [ZLO+21]), where N_{vf} can take on the values 1, 2 or 3, so that the total finger width becomes $W_{f,tot} = N_{vf} W_f$. For a given N_{vf}, one has $N_{vf} + 1$ gate contacts, so that increasing N_{vf} results in lower gate resistance, with beneficial effects on the f_{max} of the transistor, as discussed

in Chapter 3. The gate length is $L_g = N \times L_{g,ref}$, where $L_{g,ref}$ is the smallest value, once again different compared to 16FF and N can take on the values 1, 1.11, 1.33, 1.56 and larger.
In planar technologies, including 22SOI, the conducting channel is a planar sheet located below the gate finger (Figure 2.11a), so that the total active width of the transistor is given by:

$$W_{tot} = W_f N_{fing} M \tag{2.8}$$

Conversely in FinFET technologies the conducting channel is formed on the outer surface of the three-dimensional fins (Figure 2.11b), so that the device effective width is equal to:

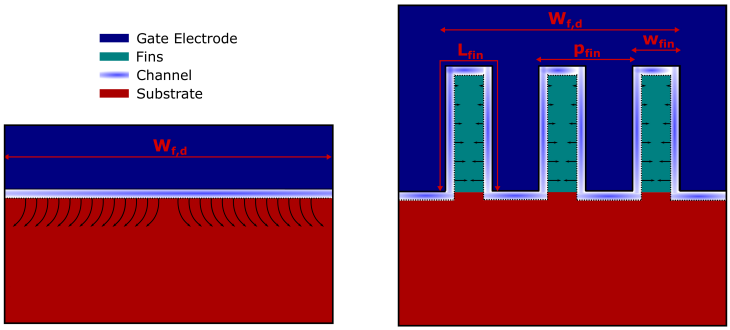
$$W_{tot,eff} = N_{fins} L_{fin} N_{fing} M \tag{2.9}$$

where L_{fin} is the total fin length. One can also define a width based on the physical size of the gate finger, as in (2.8), which goes under the name of drawn width:

$$W_{tot,d} = (w_{fin} + p_{fin}(N_{fins} - 1)) N_{fing} M \tag{2.10}$$

where p_{fin} is the fin pitch and w_{fin} is the width of a single fin. While in a planar technology $W_{tot,d}$ and $W_{tot,eff}$ coincide, in FinFET they can be significantly different. In 16FF one has $L_{fin} \sim 2p_{fin}$, so that $W_{eff} \sim 2W_d$, which has severe consequences on the reliability of the device (see section 4.4.2).

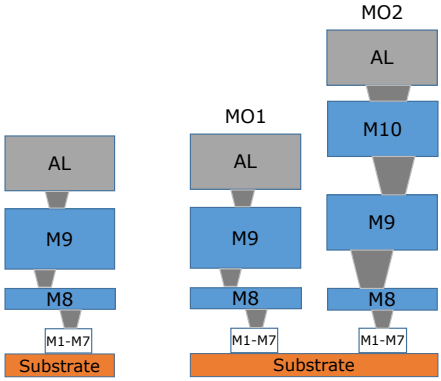
In 16FF the RF metal stack consists of 9 copper layers (Mi) + 1 aluminum layer (AL), as shown in Figure 2.12a. The 22SOI process instead offers two different metal stack options, one almost identical to 16FF (9Mi + 1AL) called MO1 and one with an additional copper thick layer (10Mi + 1AL) called MO2 (Figure 2.12b). One important technological question tackled in Chapters 4 and 5 of this thesis is whether or not the more expensive MO2 results into significant



(a) (b)

Figure 2.11: Illustration of the geometrical features of a single gate finger in (a) a planar and (b) a FinFET transistor.

performance improvement over MO1. Finally, due to the reasons explained in section 2.2.3, both 16FF and 22SOI have a ρ_{sub} of only a few Ω cm.



(a) (b)

Figure 2.12: Qualitative representation of the RF metal stacks in (a) 16FF and (b) 22SOI.

2.3 Design Considerations

Designing mmW circuits and specifically PAs is in general a challenging task regardless of the utilized technology because at such high frequencies the gain of the active devices degrades and the losses of the interconnects become significant. This calls for special attention in the layout phase, which requires careful extraction and simulation of the parasitics. As explained at length in Chapter 5, an additional complication comes into play if deeply scaled CMOS technologies are used, as the increased Impedance Transformation Ratio (ITR) required by the output MNW results in additional degradation. In the last few decades this problem has been solved using circuit architectures based on device stacking [KK15, DHG+13] and power combining [NCC12, ZR14, ALK+08]. Device stacking consists in connecting N identical active devices as shown in Figure 2.13a in such a way that the supply voltage can be increased by a factor N with respect to a single CS stage.

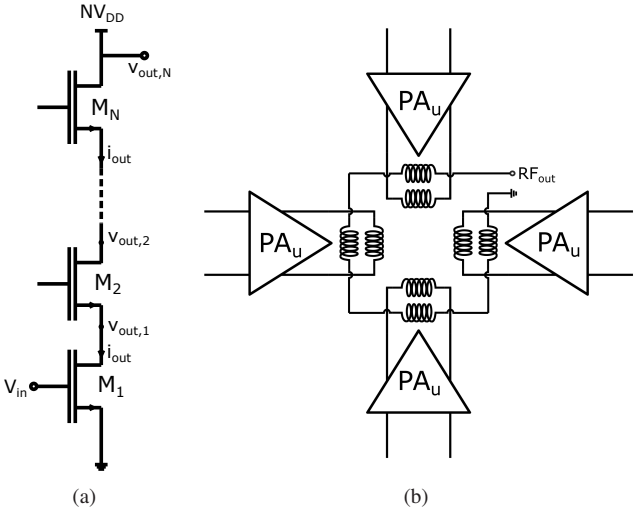


Figure 2.13: Illustration of the (a) FET stacking and (b) power-combining PA architectures.

Designing the circuit in such a way that the supply voltage and the voltage swing are equally distributed among the N devices, the output voltage at the ith drain node can be written as:

$$V_{out,i}(t) = iV_{DD}(1 + \sin(\omega_0 t)) \tag{2.11}$$

Denoting $|v_{out}|$ and $|i_{out}|$ the voltage and current amplitudes in a single CS stage, (2.11) shows that the voltage amplitude at the top drain of the N-stacked device is $|v_{out,N}| = N|v_{out}|$. Moreover the circuit topology constraints the current signal amplitude to be $|i_{out,N}| = |i_{out}|$. It follows that $R'_{L,opt} = NR_{L,opt}$, where $R'_{L,opt}$ is the optimum load resistance of the N-stacked device and $R_{L,opt}$ that of a single CS stage. Assuming that the voltage swings across the various devices are perfectly phase-aligned and that the load impedance is purely real, the theoretical output power of the N-stacked device is given by:

$$P_{out,N} = \frac{NV_{DD}^2}{2R_{L,opt}} = NP_{out} \tag{2.12}$$

where P_{out} is the output power of a single CS stage (2.2). Stacking reduces also the ITR required by the MNW, thus providing a twofold benefit.

Power combining consists in summing up the output power of several identical unit PAs (PA_u) by means of a passive power combiner. Figure 2.13b shows an example with four differential unit PAs and a four-way transformer-based series combiner. Since the active size of each unit PA is a quarter of that which would be required to obtain the same P_{out} from a single PA, $Z_{L,opt}$ is also larger, which results once again in a relaxation of the ITR of the output MNW. Thanks to the the extensive usage of device stacking and power combining, a large number of mmW PAs in deeply scaled CMOS technologies with excellent performance have been demonstrated. Designs of this type have emerged in all the CMOS flavors mentioned in section 2.2.3, namely planar bulk [TNH14, SPD+16, DDT+20], planar SOI [CHA+13, AJA+14] and bulk FinFET [DDT+20, CCW+21]. The work in [DDT+20] is particularly interesting because it shows how the stacked architecture can be easily extended to a differential configuration and utilized as building block within a power combining concept (see Figure 2.14).

It should be considered that, as helpful as they might be, device stacking and power combining are not free of issues. Stacked devices typically suffer from phase misalignment among the drain voltage signals [DHG+13] and reliability issues, whereas power combining architectures suffer from large area consumption and power combiner loss [ZR14]. This translates into a limit on the

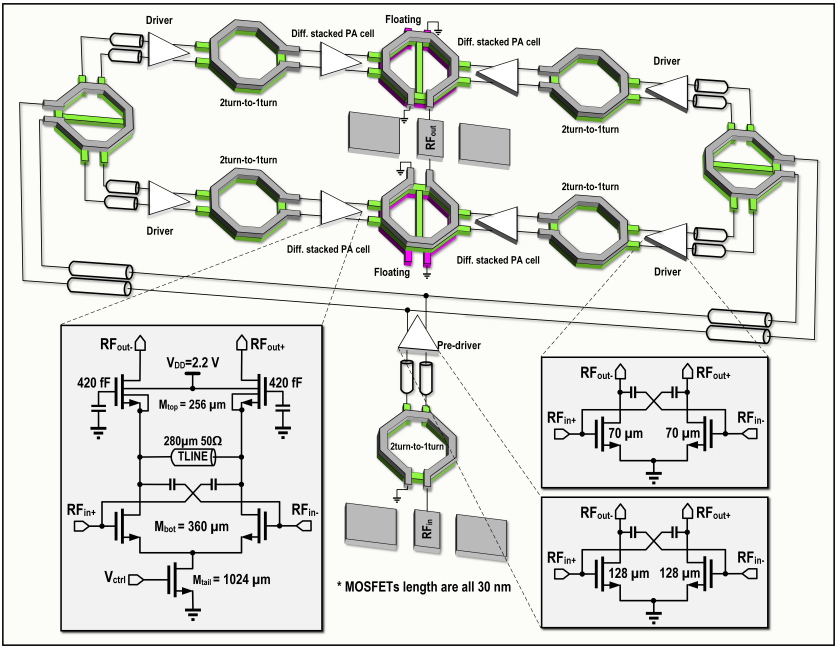


Figure 2.14: Schematic of the 28nm bulk CMOS PA from [DDT+20], featuring extensive usage of device stacking and power combining (reprinted from [DDT+20] ©2020 IEEE).

maximum number of devices which can be stacked, or unit PAs which can be power-combined, without incurring excessive efficiency degradation. The scope of this thesis is not to propose a novel circuit design concept, but rather to investigate the fundamental limitations of a technology and its suitability for the design of mmW PAs. A critical part of this effort is to investigate at length the trade-off between the active size of the PA and the insertion loss of the MNW. For this reason, stacking and power combining are deliberately not used throughout this work.

3 Gate Resistance Characterization Techniques

3.1 Introduction

The small-signal model of a FET in CS configuration operating in the saturation region, shown in Figure 3.1, is one of the foundations of analog circuit design. It allows to express the key RF figures of merit such as f_t and f_{max} in terms of some circuit components which depend on the physical properties of the transistor. It includes the parasitic gate (R_g), source (R_s) and drain (R_d) resistances, the parasitic gate-to-source (C_{gs}), gate-to-drain (C_{gd}) and drain-to-source (C_{ds}) capacitances and the output resistance r_0 to model the channel length modulation effect. It also includes a voltage-controlled current source to model the drain current $i_d = g_m v_{gs}$, where g_m is the transconductance of the transistor and v_{gs} is the gate-to-source voltage signal.

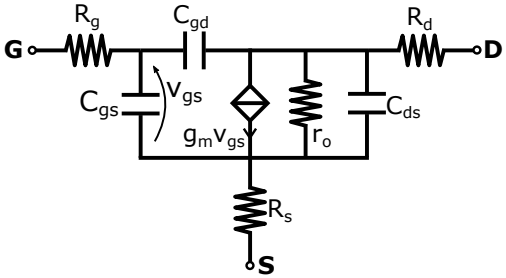


Figure 3.1: Small-signal model of a CS FET in saturation.

From this circuit model one easily obtains [GP17]:

$$f_t = \frac{g_m}{2\pi \left(C_{gs} + C_{gd}\left(1 + g_m(R_s + R_d)\right)\right)} \tag{3.1}$$

$$f_{max} = \frac{f_t}{2\sqrt{\frac{g_m C_{gd}\left(R_g + R_s\right)}{C_{gs} + C_{gd}} + \frac{R_g + R_s}{r_o}}} \tag{3.2}$$

The equations above show that the gate resistance does not affect the f_t of the transistor but it does have a significant impact on f_{max} [Lit01]. It also has a major impact on the noise figure NF [RYL94], which is however not relevant for this work. From the analog circuit perspective, this translates into a severe limitation mainly for mmW front-end blocks such as PAs and LNAs. In order to keep R_g low, circuit designers have to select a suitable geometry for the active devices [NCC12], and to do so it is crucial that the behavior of R_g be correctly captured in the compact model of the transistors. In the last years a lot of research work has been published on this topic, achieving very good results [JOC+98, KKJS07, CM01, DSC+12, CTC+14]. Since the target of any model is to reproduce measurements as closely as possible, using the best-known measurement methodology is a fundamental pre-requisite.

One key aspect in device characterization at mmW frequencies is the de-embedding of the on-chip interconnect parasitics, which are caused mainly by the measurement pads and by the feedlines connecting the Device Under Test (DUT) to the pads themselves. Most of the traditional de-embedding methods are based on a lumped representation of the parasitics and make use of one or more auxiliary structures to eliminate their contributions and extract the behavior of the DUT. The most popular is the open-short method [KGV91], which is used as reference in this Chapter. The main limitation of the lumped methods is that they neglect the distributed nature of the interconnects, which results in degrading accuracy as the frequency increases. In order to partially take this effect into account, some more refined methods have been devised [KCL12, THJB05], which achieve higher accuracy but require additional de-embedding structures and therefore larger chip area.

In the analyzed literature the standard structure with a single transistor in CS configuration and open-short de-embedding is consistently used for the extraction of R_g. In this work we consider also an alternative structure with the transistor connected in capacitor mode, called "capacitor-like" structure, which requires only the open de-embedding step. The Chapter is organized as follows: in section 3.2 the physical origin of the various contributions of the gate resistance is briefly explained. In section 3.3 the main features of the

two measurement methodologies are presented along with a list of fabricated test structures in the 16FF process. In section 3.4 the measurement setup is described and some of the figures of merit utilized throughout the Chapter are introduced. In section 3.5 the capacitor-like structure is analyzed in detail and some design guidelines are derived to achieve accurate measurement results. In section 3.6 the standard and capacitor-like structures are compared and finally in section 3.7 the conclusions of this study are summarized. The main results presented in this Chapter are extracted from [4].

3.2 Physical origin and modeling

Over the years the gate stack of the MOS transistor has undergone several developments, mainly driven by the need to maintain good electrostatic control of the channel and gate isolation in spite of the continuous scaling. A major breakthrough was achieved in the first decade of the 2000s with the transition from the SiO_2 oxide layer with polysilicon gate electrode to the high-k dielectrics with metal gate electrode [DAB03, RW15]. Although this was mostly aimed at improving the digital performance, significant benefit was observed also in the analog domain thanks to the lower resistivity of the metal gate compared to the polysilicon one, which resulted in lower R_g [VRC+11]. From the technological standpoint, the change of gate electrode material posed new challenges to the traditional Gate First (GF) fabrication process, which exploits the polysilicon metal gate for the self-aligned implantation of the source and drain extensions. This is possible because the polysilicon, unlike most metals, can withstand the high temperatures required by the annealing step after the dopant implantation. A common solution was the adoption of the so-called Gate Last (GL) process, in which the gate metal deposition is performed at the end [Fra11]. In order to maintain the self-alignment capability offered by the polysilicon, the Replacement Metal Gate (RMG) technique was devised, in which a dummy polysilicon gate is used for the dopant implantation and subsequently replaced with a metal. For quite some time, the GF and GL approaches coexisted, leading to two different gate stack structures. The GL stack, shown in Figure 3.2b, consists of a pure metal gate on top of a layer of a so-called Work Function (WF) metal, which is required to obtain the desired threshold voltage. On the other hand the GF stack, shown in Figure 3.2a, consists of a polysilicon layer on top of a thin WF metal layer. Besides defining the threshold voltage, in this case the WF

metal reduces the gate depletion of the polysilicon gate, which would otherwise lead to an undesired increase of the electrical oxide thickness. Furthermore, it prevents reactions between the high-k dielectric and the polysilicon at high temperatures, ensuring the stability and performance of the device. Finally, a silicide layer is also added on top of the stack to lower the gate resistance. To this day, most foundries have transitioned to the GL approach for deeply scaled nodes.

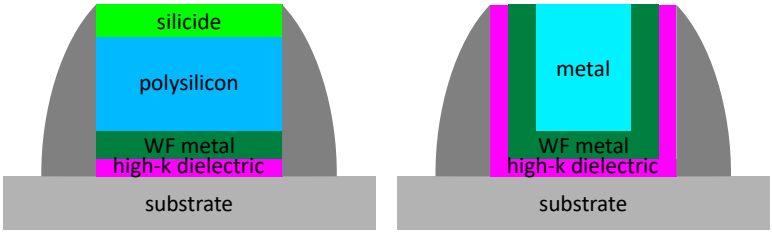


Figure 3.2: Schematic representation of the (a) GF and (b) GL stacks.

The modeling of R_g is a rather complicated topic due to the multiple contributions involved and the non-trivial dependency on the transistor geometrical parameters. In CMOS technologies a multi-finger layout is normally used for RF transistors to achieve the desired device width while keeping R_g low. Considering the relatively simple case of a planar technology, the input resistance presented by a single gate finger can be modeled by means of a distributed Resistive-Capacitive (RC) network, as shown in Figure 3.3. The main contributions are the bias-independent electrode resistance, which can be decomposed into a horizontal ($r_{el,h}$) and a vertical component ($r_{el,v}$), and the bias-dependent channel resistance $r_{ch}(V_{gs})$, which is connected to the electrode through the oxide capacitance c_{ox} [JOC+98, KKJS07, LGC+97, DSC+12]. The relative magnitude of these components depends strongly on the gate process. Indeed GF typically shows lower horizontal component due to the silicides and larger vertical component due to the interfaces between silicides and polysilicon and between polysilicon and WF metal. On the other hand GL shows larger horizontal component due to the higher resistivity of the WF metal but lower vertical component due to the absence of the silicide layer. Moreover, since the WF metal extends along the vertical sides of the gate electrode, for short gate length a significant lateral resistance component arises as an effect of the

reduced metal volume.

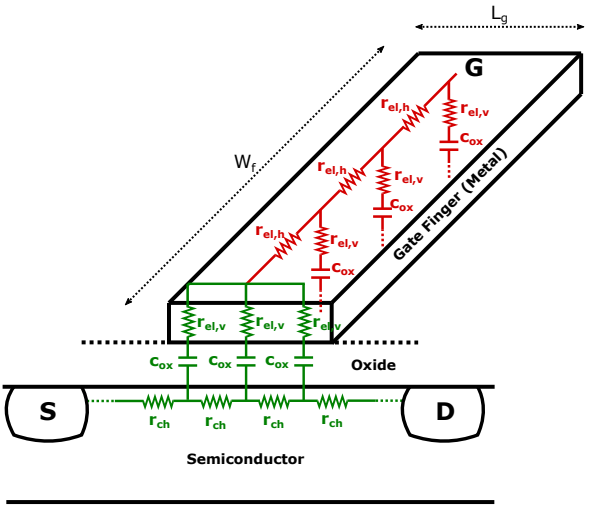


Figure 3.3: Distributed model of a single gate finger of a MOS transistor, including the gate electrode resistance, the oxide capacitance and the channel resistance contributions.

Unfortunately this distributed representation of the MOS structure can not be easily integrated in a compact model of the transistor, where it is preferable to use only one node for each terminal to prevent long simulation times at circuit level. For this reason the distributed network is typically simplified into a single lumped resistance R_g placed at the input of the equivalent circuit [EC00] as in Figure 3.1, which can be written in the form:

$$R_g = R_{el,h} + R_{el,v} + R_{ch} \tag{3.3}$$

where $R_{el,h}$, $R_{el,v}$ and R_{ch} are lumped equivalent resistances which absorb the contributions of the elementary resistances of Figure 3.3. In terms of the device parameters, $R_{el,h}$ is proportional to W_f/L_g, whereas $R_{el,v}$ is proportional to L_g/W_f [KKS05], where W_f is the finger width and L_g the gate length. Since each component of the model has to be bias-independent, the dependency of R_{ch} on V_{gs} is normally sacrificed, and the value at one typical V_{gs} operating point is chosen.

In FinFET technologies the structure of the gate electrode is more complicated than in planar due to the inhomogeneous profile of the gate finger resulting from the presence of the fins [WC05, MPR+15]. This gives rise to additional components of R_g, so that careful optimization of the transistor is required to limit R_g to acceptable values. This is one of the reasons behind the choice of the 16FF process for the investigation carried out in this Chapter.

3.3 Measurement structures for the gate resistance

The standard method for the measurement of the gate resistance is based on the structures in Figure 3.4. The main structure in Figure 3.4a consists of an RF transistor in CS configuration, the so-called DUT, routed to RF Ground-Signal-Ground (GSG) pads. The open in Figure 3.4b is obtained from the main structure removing the DUT, whereas the short in Figure 3.4c is obtained from the open shorting the input and output feedlines to ground. These are the structures which are commonly used to extract R_g as well as the other equivalent-circuit parameters of the transistor. Using the small-signal equivalent circuit in Figure 3.1, the gate resistance can be extracted from the 2-port Y-parameters using the formula $R_g = \mathrm{Re}(1/Y_{11})$ [DSC+12], under the assumption that the source and drain parasitic resistances R_s and R_d are negligible with respect to R_g.

The alternative capacitor-like structure in Figure 3.5 consists of an RF transistor with the gate connected to both the input and output pads, and source and drain shorted to ground. The naming is due to the fact that in this configuration the channel is shunted out and the transistor behaves as a capacitor C_{gg} in series with R_g, where $C_{gg} = C_{gs} + C_{gd}$ is the total gate capacitance of the transistor.

One key advantage of the capacitor-like structure is that it requires only the open de-embedding step. Indeed, once the shunt parasitic components introduced by the pads are removed with the open de-embedding, one is left with a T-network formed by the feedlines and the DUT, as shown in Figure 3.6. Taking Z_{21} of this network automatically excludes the contribution of the feedlines (Z_{fl}) and no additional de-embedding step is required. Based on considerations very similar to those done for the standard structure, it is found that $R_g = \mathrm{Re}(Z_{21})$, again under the assumption that $R_s, R_d \ll R_g$. It should be noted that this concept can not be used in 1-port configuration, as it would require both the open and

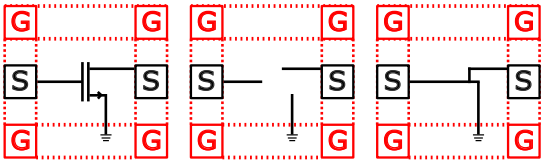


Figure 3.4: Standard structures for Rg measurement: (a) Main, (b) Open and (c) Short.

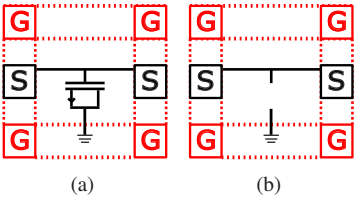


Figure 3.5: Capacitor-like structures for Rg measurement: (a) Main and (b) Open.

short de-embedding steps.

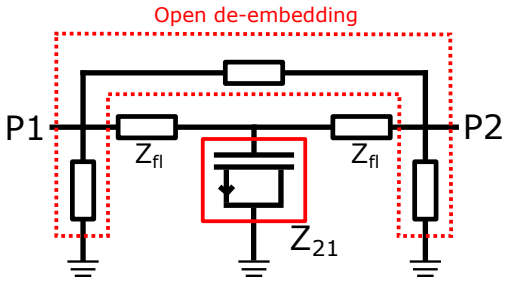


Figure 3.6: Illustration of the de-embedding methodology in the capacitor-like structure.

For this study 18 test structures utilizing both the standard and capacitor-like concept were fabricated in the 16FF process. The list of all the available structures with the corresponding geometrical features is presented in Table 3.1. All the utilized devices are RF transistors with the lowest threshold voltage available in the PDK, approximately 0.2 V. The various structures differ in the values of the transistor parameters that have the strongest impact on R_g, namely

N_{fins}, L_g and M. On the contrary N_{fing} and PP are fixed to 10 and $1.07 \times PP_{ref}$ respectively because of their weaker influence. In addition to several instances of the capacitor-like structure, three standard structures with different values of M (1,4,8) were fabricated for comparison. For both types of structures, the on-chip interconnections are de-embedded up to the third level of metallization (M3).

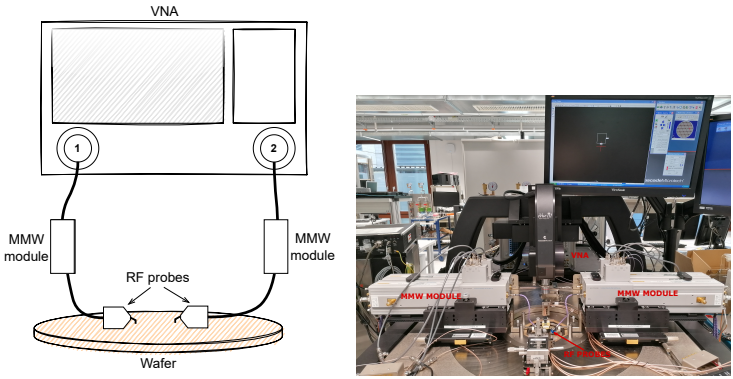
Table 3.1: List of Gate Resistance test structures on the 16FF testchip.

DUT	Structure Type	N_{fins}	L_g (nm)	M
1	Capacitor-like	6	20	1
2	Capacitor-like	10	20	1
3	Capacitor-like	16	20	1
4	Capacitor-like	20	20	1
5	Capacitor-like	6	20	4
6	Capacitor-like	10	20	4
7	Capacitor-like	16	20	4
8	Capacitor-like	20	20	4
9	Capacitor-like	6	20	8
10	Capacitor-like	10	20	8
11	Capacitor-like	16	20	8
12	Capacitor-like	20	20	8
13	Capacitor-like	20	16	8
14	Capacitor-like	20	18	8
15	Capacitor-like	20	24	8
16	Standard	20	20	1
17	Standard	20	20	4
18	Standard	20	20	8

3.4 Measurement setup, simulation setup and figures of merit

The measurement setup consists of a FormFactor Elite 300/AP-0011 Probe Station for 300 mm wafers, a Keysight N5227A PNA Vector Network Analyzer (VNA) with frequency range from DC up to 67 GHz and FormFactor Infinity RF GSG probes for on-wafer probing with frequency range from DC up to

110 GHz. In order to be able to measure the S-parameters up to 110 GHz, the frequency range of the VNA is extended using a Keysight N5250CX10 mmW module for each port. The output of the mmW module and the input of the RF probes are connected using 1-mm coaxial cables. For the 2-port S-parameter measurement the VNA is calibrated up to the probe tips using a standard 2-port Short-Open-Load-Thru (SOLT) method. A block-diagram and a photograph of the measurement setup are shown in Figure 3.7a and 3.7b respectively.



(a) (b)

Figure 3.7: (a) Block-diagram and (b) photograph of the measurement setup.

The measurements were carried out with RF input power $P_{in} = -20\,\text{dBm}$ at both ports, for values of the gate bias voltage V_{gs} ranging between 0 V and 0.8 V. For the standard structure, a drain bias voltage $V_{DD} = 0.8\,\text{V}$ was used. In order to assess the quality of the measurement, the relative deviation ΔR_g of the measured gate resistance ($R_{g,meas}$) from the simulated one ($R_{g,sim}$) was used:

$$\Delta R_g = \frac{R_{g,meas} - R_{g,sim}}{R_{g,meas}} \tag{3.4}$$

One issue with this figure of merit is that the details of the device model from the foundry are not known. For this reason we first verified that $R_{g,sim}$ follows the expected scaling law with respect to N_{fins} and M [HN19], given by:

$$R_g = \frac{R_{conn} + R_{el,v}/N_{fins} + R_{el,h} \times N_{fins}}{M} \tag{3.5}$$

where R_{conn} is a constant which includes end resistances, contact resistances and interconnects up to M3. This result, shown in Figure 3.8, justifies the usage of the foundry model as reference to assess the quality of the measured data.

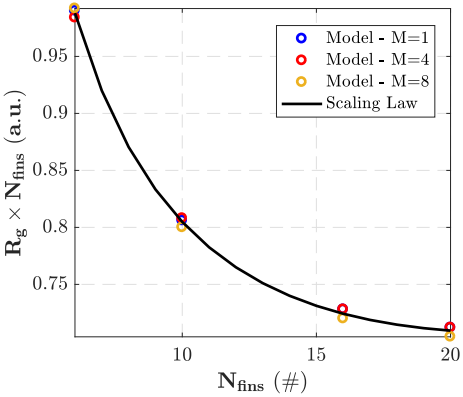


Figure 3.8: Scaling behavior of R_g vs N_{fins} and M obtained from the foundry model and from the scaling law (3.5) with $R_{el,v} = 155.4\,\Omega$, $R_{el,h} = 0.3\,\Omega$, $R_{conn} = 21.7\,\Omega$ for an RF transistor with $N_{fing} = 10$, $V_{gs} = 0.4\,V$ at $f_0 = 50\,GHz$.

3.5 Capacitor-like structures

This section focuses on the analysis of the capacitor-like structure. The plot of R_g vs frequency for DUT12 with $V_{gs} = 0.4\,V$ in Figure 3.9a shows very good agreement with the foundry model over the entire frequency range. On the other hand the plot of $|\Delta R_g|$ over frequency for different bias conditions in Figure 3.9b shows that the best agreement between measurement and simulation is obtained for $V_{gs} = 0.4\,V$. The reason is that, as explained in section 3.2, the dependency of R_{ch} on V_{gs} is neglected in the model, and a single value at a "convenient" V_{gs} is taken. Based on these data, the chosen bias point seems to be $V_{gs} = 0.4\,V$, which is a reasonable choice, as it was observed to be the bias condition which optimizes f_t. In practice, since the transistor is biased most of the times close

to V_{gs} = 0.4 V, the error introduced by this approximation is small.

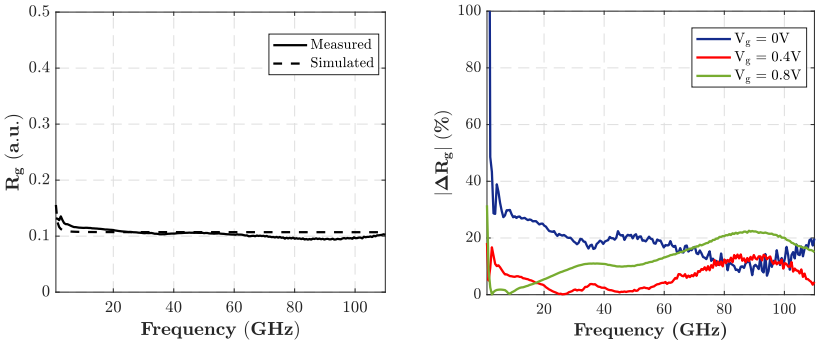


Figure 3.9: (a) Measured and simulated R_g at V_{gs} = 0.4 V and (b) $|\Delta R_g|$ for different values of V_{gs}. Both quantites refer to DUT12 and are plotted as a function of frequency.

Table 3.2 reports the values of ΔR_g for different DUTs at V_{gs} = 0.4 V and f_0 = 50 GHz. The value f_0 = 50 GHz is chosen because it is approximately in the middle of the analyzed frequency range. It can be observed that a minimum total device width is required to achieve good agreement between measurement and simulation. The reason is that for the smallest devices like DUT1, the total gate capacitance C_{gg} of the transistor is smaller or comparable to the pad capacitance $C_{pad} \sim 25$ fF, which results in a large numerical error in the open de-embedding step. Based on these considerations, a large value of M should be used if the width of the transistor is small. This is not necessary if the width of the device is large enough, as in the case of N_{fins} = 20.

Table 3.2: ΔRg in % at f_0 = 50 GHz with V_{gs} = 0.4 V for capacitor-like structures using transistors with various combinations of N_{fins} and M.

M	**N_{fins}**			
	6	**10**	**16**	**20**
1	-52	-18.3	-5.6	3.9
4	9	3.8	4.2	-1.6
8	1.7	-1.8	-4.2	-1.7

In Figure 3.10 the measured R_g is compared to simulations as a function of

the geometrical parameters N_{fins} and L_g. In order to capture the device-to-device (mismatch) variations and the die-to-die, wafer-to-wafer and lot-to-lot (process) variations, Monte Carlo simulations with 2000 samples are run for each set of parameters. The results are displayed in 3 different curves: the mean value $\mu(R_g)$ of the gate resistance and the so-called $\pm 3\sigma$ curves, i.e. the quantities $\mu(R_g) \pm 3\sigma(R_g)$, where $\sigma(R_g)$ is the variance of R_g. These are relevant because they define the interval in which R_g falls with a probability of 99.7%, and therefore provide a good estimation of the process and mismatch variation. The measured data show very good correlation with $\mu(R_g)$ and lie completely in the interval delimited by the $\pm 3\sigma$ curves. Based on these simulation results, an overall R_g fluctuation of up to 55% above or below the mean value is expected. One interesting observation from Figure 3.10b is that the spread of R_g becomes tighter for large values of L_g. This is expected because L_g is one of the transistor parameters which is most affected by the process variation. Since the fluctuation δL_g is independent of L_g, its impact decreases as L_g gets larger.

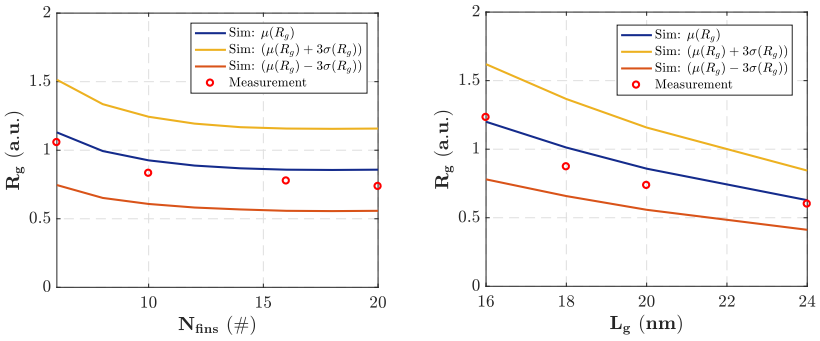


Figure 3.10: Measured and simulated R_g from capacitor-like structures with $M = 4$ and $V_{gs} = 0.4\,V$ at $f_0 = 50\,GHz$ (a) as a function of N_{fins} (with $L_g = 20\,nm$, $N_{fing} = 10$) and (b) as a function L_g (with $N_{fins} = 20$, $N_{fing} = 10$).

3.6 Comparison between standard and capacitor-like structures

In order to make an effective comparison between the two types of structure, standard DUTs 16, 17 and 18 have been included in the testchip, having the same transistor parameters as capacitor-like DUTs 4, 8 and 12 respectively. Comparing ΔR_g of the 3 pairs of structures, it is found that the standard structure gives the best results for M = 1, as shown in Table 3.3. Larger values of M (4 and 8) lead to larger deviations and should be avoided. In this specific case the capacitor-like structure is not very sensitive to M due to the large device width (N_{fins} = 20), but in general it shows the opposite behavior, as discussed in section 3.5. Differently from the capacitor-like structure, in the standard structure R_g is not connected in parallel with C_{pad}, therefore the higher C_{gg} resulting from the larger M does not bring any advantage to the measurement. On the contrary, the larger number of devices in parallel exacerbates the error caused by the two-step de-embedding methodology. Therefore the recommendation is to keep M as low as possible, which is exactly the opposite as in the case of the capacitor-like structure. All in all, the achievable ΔR_g with the two structures is comparable if the recommended value of M is used in each case.

Table 3.3: ΔRg in % at f_0 = 50 GHz with V_{gs} = 0.4 V for standard and capacitor-like structures with N_{fins} = 20, N_{fing} = 10, L_g = 20 nm and different values of M.

Structure Type	**M**		
	1	**4**	**8**
Standard	1.7	16.4	21
Capacitor-like	3.9	-1.6	-1.7

The second important comparison criterion is the variation of the measured R_g over frequency, which could be potentially influenced by the de-embedding structures. It can be quantified by means of the normalized standard deviation over frequency $\hat{\sigma}_{R_g}/\overline{R_g}$, where $\overline{R_g}$ and $\hat{\sigma}_{R_g}$ are respectively the mean value and the standard deviation of R_g over frequency, defined as:

$$\overline{R_g} = \frac{1}{N}\sum_{i=1}^{N} R_g(f_i) \tag{3.6}$$

$$\hat{\sigma}_{R_g} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_g(f_i) - \overline{R_g})^2} \tag{3.7}$$

with N being the number of frequency points. The normalized standard deviation is plotted in Figure 3.11 as a function of V_{gs} for the standard and capacitor-like structures with different values of M. It can be observed that the measurements performed with the two structures show a variation over frequency between 2% and 10%. In most cases the variation is between 2% and 5%, with the exception of the capacitor-like structure with M = 1 and the standard structure with M = 8, which show up to 7% and 10% deviation respectively. This shows that the standard structure with large M and the capacitor-like structure with small M represent the worst case not only in terms of agreement with the model, as discussed in sections 3.5 and 3.6, but also in terms of variation over frequency.

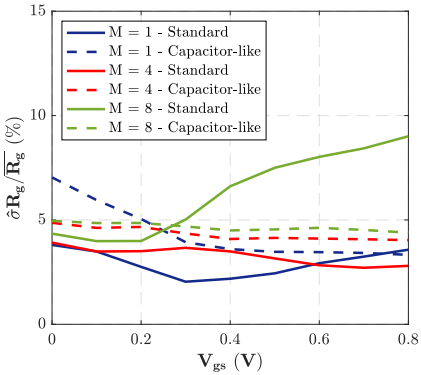


Figure 3.11: Normalized variance of R_g over frequency as a function of V_{gs} for standard (DUTs 16,17,18) and capacitor-like (DUTs 4,8,12) structures.

3.7 Summary

This chapter provided an overview of the physical origin of the gate resistance in MOS transistors and discussed its impact on the performance of RF and mmW

circuits. It presented two different methodologies for the characterization of the gate resistance itself, based on the standard and capacitor-like structures, which were analyzed and compared with the aid of test structures in the 16FF process. It was found that the design guidelines to achieve best accuracy in the two types of structure are somehow opposite: for the standard structure there is a constraint on the maximum transistor size, whereas for the capacitor-like structure on the minimum size. Following these guidelines, the two methods achieve similar agreement with the foundry model and similar variation over frequency. For the capacitor-like structure, a process and mismatch variation study based on Monte Carlo simulations was performed, showing that the gate resistance can fluctuate up to 55% above and below the average value.

4 Design of the Amplifying Stages

Nearly all published mmW PAs are designed for a specific application using a semiconductor technology which allows to meet the desired specifications. Drawing conclusions on the performance of different technologies with respect to one another based on this kind of work is typically cumbersome because they are all based on different architectural solutions and on the optimization of different parameters. Substantial literature has been published specifically on technology comparison topics [JJO+18, ARÖ+19], but most of it is limited to the comparison of standard figures of merit like f_t, f_{max}, BV_{off} and Noise Figure (NF). The main contribution of this work is to establish a systematic technology benchmarking methodology specifically tailored for mmW PAs, which takes into account the large-signal performance. This Chapter deals with the design of the amplifying stages of the PA, which are based on a simple circuit architecture widely used in CMOS technologies, as explained in section 4.1. The center frequency of the PA is chosen to be $f_0 = 80\,\text{GHz}$ because it is sufficiently large for the amplifying stage to be significantly affected by the layout parasitics. Moreover it is located in the E-band, which is of interest for many applications such as automotive radar (see section 1.3.2) and mobile backhaul. In section 4.2 some application-independent design criteria are chosen, which serve as foundation for an algorithmic design methodology. In section 4.3 this methodology is applied to the design of two PA output stages with optimum performance in the 16FF and 22SOI processes respectively. In 22SOI the impact of the metal stack on the performance of the amplifying stage is analyzed comparing the two available stack profiles for RF applications. Finally in section 4.4 the electromigration phenomenon and its different relevance in planar and FinFET processes is analyzed, using the designed 16FF and 22SOI amplifying stages as test vehicles. The design of the MNWs is also a fundamental part of this methodology, but it is tackled separately in Chapter 5 due to the different optimization criteria. The main results related to 16FF presented in sections 4.1 through 4.3 are extracted from [3].

4.1 Circuit Topology, Layout and Extraction Methodology

In order to enable a fair comparison among different technologies, a relatively simple and widespread circuit architecture utilizing the Neutralized Differential Pair (NDP) [AMB+11, DN10, ZR13] (Figure 4.1a) is used for the amplifying stages. These are built out of a NDP unit device with a compact layout concept very similar to that proposed in [CPH19, YGH+22], as shown in Figure 4.1b. Each unit device consists of two identical transistors and two cross-coupled Metal-Oxide-Metal (MOM) neutralization capacitors C_n, all from the PDK. The selected active devices are those with RF-optimized layout and with the lowest threshold voltage, so as to maximize the output voltage and current swings. The value of C_n is selected in such a way to attain perfect neutralization, that is maximum stability factor K_f.

The PA stage is particularly sensitive to the gate and drain parasitic resistances (R_g, R_d), which cause degradation of G and P_{out} respectively. In order to minimize these resistances, the gate and drain terminals of the unit cell are routed to the uppermost copper layer of the metal stack (M_{top}), using the largest possible trace widths and number of vias allowed by Design Rule Checking (DRC). In an earlier version of the design, the gate was routed only up to M_{top-1} to limit C_{gd}. However subsequent iterations revealed that extending this routing up to M_{top} provides a better outcome, as C_{gd} can be fully compensated by C_n regardless of its magnitude. In order to minimize the imbalance effects between the two differential terminals, the highest possible symmetry about the horizontal axis is maintained in the layout of the unit cell. The only exception are the neutralization paths, which have to be drawn on opposite sides of the active device due to layout constraints.

In order to achieve the desired P_{out}, the utilized layout concept allows to connect a number M of unit cells in parallel side-by-side in a seemless way, as shown in Figure 4.1c. The utilized concept with local neutralization in the unit device has been preferred over the one with only two neutralizing capacitors for the entire stage [ZR14]. This solution improves the modularity of the design making C_n less sensitive to the interconnect parasitics and leads to a larger quality factor Q_n of the neutralization path. Moreover the placement of the capacitors on the upper and lower side of the unit cell does not cause any overhead in the layout parasitics. Large values of M, which are typically required for the PA core, give rise to long gate (RF_{inp}, RF_{inm}) and drain (RF_{outp}, RF_{outm}) lines,

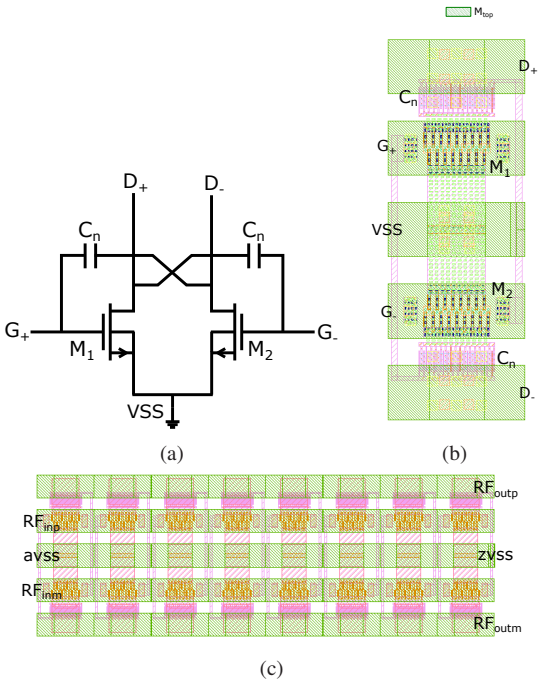


Figure 4.1: NDP-based PA amplifying stage: (a) circuit schematic, (b) layout of a unit cell and (c) layout of a full stage with M = 8.

which show substantial transmission-line behavior at the design frequency. As M increases, the output signals of the different unit devices reach the output of the stage with slightly different attenuations and phases, which translates in a degradation of the Maximum Available Gain (MAG). In the case of the output stage, this poses a limit to the maximum W_{tot} which can be used without excessive performance loss. This effect is shown in Figure 4.2 for different parasitic extraction methodologies, namely RC extraction, RLCk extraction and EM simulation. The fact that RLCk predicts a stronger degradation than RC for $M \geq 8$ confirms that the parasitic inductance plays a critical role at this frequency and cannot be neglected. Moreover, since the EM simulation predicts an even stronger degradation than RLCk for large M and it is known

to be more accurate for the extraction of inductive parasitics, it was decided to use it as standard extraction methodology in this work.

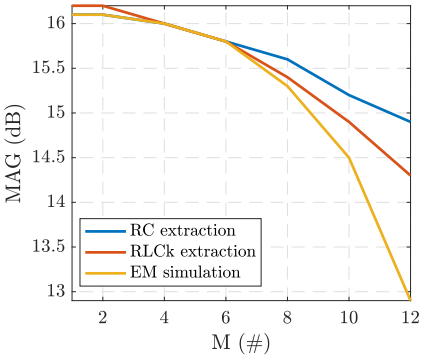


Figure 4.2: MAG of PA amplifying stage in 16FF with N_{fins} = 12, N_{fing} = 16, $L_g = 1.5 \times L_{g,ref}$, $PP = 1.44 \times PP_{ref}$, V_{gs} = 0.5 V at 80 GHz as a function of M with different parasitic extraction methodologies.

4.2 Design Criteria

Depending on the application, the PA has to generate a given output power P_{out}, which requires a certain total active device width W_{tot} for the output stage. Unfortunately, as W_{tot} gets larger, the increased interconnect parasitics and ITR of the output MNW cause significant degradation of G and, consequently, of PAE. As mentioned in section 2.3, these issues can be partially overcome using the stacking and power-combining architectures. However, since one goal of this work is to investigate the capabilities and limitations of a certain technology, these techniques are deliberately not used. For the same reason, as mentioned in the introduction to this Chapter, this study does not aim at meeting the specifications of a given application, but targets a joint optimization of the technology-dependent quantities like P_{out}, G and PAE. On the contrary, application-dependent metrics such as linearity and bandwidth are not considered. The ITRS FoM introduced in section 2.1.4 is therefore a good metric to drive the design of the output stage. One major issue is that G_{ss}, P_{sat} and PAE_{peak} correspond to different drive levels of the amplifying stage and can

not be easily maximized simultaneously. For the purpose of layout optimization, the following modified version evaluated at the 3dB-compression point is therefore adopted:

$$\mathrm{FoM_{3dB}} = \mathrm{G_{3dB}(dB)} + \mathrm{P_{out,3dB}(dBm)} + 10\log_{10}(\mathrm{PAE_{3dB}(\%)}) + 20\log_{10}(\mathrm{f_0}) \tag{4.1}$$

where the choice of the 3dB-compression point can be justified by the very same considerations presented in section 2.1.2. The performance degradation introduced by the output MNW is not included in (4.1) because it requires a separate optimization procedure, as discussed at length in Chapter 5.

4.3 Design of the PA output stages

Once the circuit architecture and the layout type are fixed as described in section 4.1, the most critical aspect of the design is the selection of the transistor parameters. As briefly mentioned in the Introduction to this Chapter, these are typically chosen based on considerations about f_t and f_{max} [ZR13, CRN09]. The novelty of the proposed methodology consists in using a well-defined figure of merit, namely $\mathrm{FoM_{3dB}}$, capable of capturing the large-signal behavior of the PA. Besides the figure of merit, the methodology requires an algorithmic design approach along the lines of [YGT+07], which is shown throughout the rest of this section for the 16FF and for 22SOI processes. It should be pointed out that $\mathrm{FoM_{3dB}}$ is strictly speaking a good figure of merit only for the output stage. This stems from the fact that P_{out} is less critical for the drivers, and it should be just large enough to provide enough input power to the subsequent stage. Nonetheless the same NDP unit device designed based on $\mathrm{FoM_{3dB}}$ is used for both the core and driver stages, which only differ in the value of M, as explained in Chapter 6.

4.3.1 Design in 16FF

In 16FF the main design parameters are the geometrical features of the active devices, such as the number of fins N_{fins}, number of fingers N_{fing}, gate length L_g, gate pitch PP and multiplicity M (see Figure 2.10a). One additional parameter is the gate-to-source bias voltage V_{gs}, which sets the operating class of the PA. The target of the design is to determine the set of parameters which maximizes FoM_{3dB} of the PA core using the layout concept shown in section 4.1. This is a challenging task due to the multidimensional nature of the problem. Moreover, for each set of parameters one has to determine the optimum complex loadpull impedance ($Z_{L,opt}$) which maximizes FoM_{3dB} (see section 2.1.2). The key strategy to simplify the task is to start from the parameters which have little or no influence on the routing parasitics, such as N_{fins}, PP, L_g and V_{gs}, as their effect can be studied by means of pre-layout simulations. Afterwards one can proceed to the optimization of N_{fing} and M, which have a major effect on the routing parasitics and therefore require post-layout simulations. From Fig. 4.3 it can be observed that FoM_{3dB} shows a peak at $V_{gs} = 0.5\,V$ and $N_{fins} = 12$. Considering that the utilized transistor has $V_t \sim 0.2\,V$, the value $V_{gs} = 0.5\,V$ corresponds to class-A operation (see section 2.1.3). The peak with respect to N_{fins} results from the interplay between P_{out} and G: while the former increases monotonically with N_{fins}, the latter follows closely the gate resistance R_g, which has an optimum with respect to N_{fins} [CM01, MPR+15].
In Table 4.1 a performance comparison for different values of PP is shown, but only for $1.07\times PP_{ref}$ and $1.44\times PP_{ref}$. The value $1\times PP_{ref}$ is not included because it results in worse peformance and does not support all the gate lengths used in this comparison. It is observed that $1.44 \times PP_{ref}$ delivers 2.4 dB larger FoM_{3dB} compared to $1.07 \times PP_{ref}$. The improvement comes from the larger spacing $\Delta_{P/S}$ between each gate (G) finger and the adjacent source (S) and drain (D) contacts, which results in lower parasitic capacitances and larger gain (Figure 4.4). Additionally, the slightly wider S/D contacts lower the associated parasitic resistances R_s and R_d, leading to higher P_{out}. Due to process-related limitations, with $PP = 1.44 \times PP_{ref}$ the only possible gate length value is $L_g = 1.5 \times L_{g,ref}$, which takes one design parameter out of the design problem.
In 16FF the PCell from the foundry offers layout options with routing up to M1 or up to M3. It was noticed that in the device with $PP = 1.44\times$ the routing up to M3 offered by the foundry is not optimal. As a matter of fact R_s and R_d can be improved significantly without violating DRC by increasing the trace

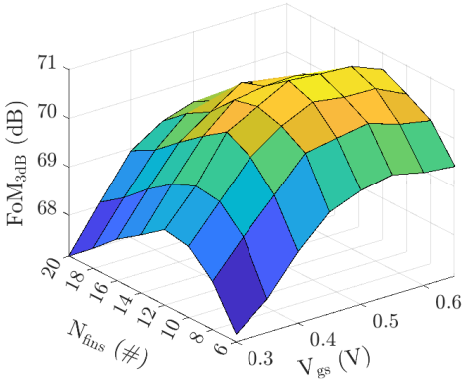


Figure 4.3: FoM_{3dB} vs V_{gs} and N_{fins} of PA core in 16FF with $N_{fing} = 16$, $L_g = 1.5 \times L_{g,ref}$, $PP = 1.44 \times PP_{ref}$, M = 1 at 80 GHz based on pre-layout simulations..

Table 4.1: Large-signal performance of PA core in 16FF with $N_{fins} = 12$, $N_{fing} = 16$, $L_g = 1.5 \times L_{g,ref}$, M = 1 at $f_0 = 80$ GHz for different values of PP (pre-layout).

PP (nm)	**V_{gs} (V)**	**$P_{out,3dB}$ (dBm)**	**G_{3dB} (dB)**	**PAE_{3dB} (%)**	**FoM_{3dB} (dB)**
$1.07 \times PP_{ref}$	0.53	6	9.9	42.8	70.3
$1.44 \times PP_{ref}$	0.53	7.1	11	45.2	72.7

widths and making the vias rectangular instead of square on the corresponding contacts. All in all, this custom routing achieves an improvement of FoM_{3dB} of about 0.6 dB compared to the standard routing, as shown in Table 4.2.

Table 4.2: Effect of M1-M3 routing on large-signal performance of 16FF PA core with $N_{fins} = 12$, $N_{fing} = 16$, $L_g = 24$ nm, $PP = 1.44 \times PP_{ref}$ and M = 1 (pre-layout) at 80 GHz.

Routing M1-M3	**$P_{out,3dB}$ (dBm)**	**G_{3dB} (dB)**	**PAE_{3dB} (%)**	**FoM_{3dB} (dB)**
Pcell Default	6.1	10.6	42.2	71
Custom	6.5	10.6	44.2	71.6

In order to determine the optimum values of N_{fing} and M a post-layout analysis is required. The layout of the unit cell is shown in Figure 4.1b, where M_{top} is

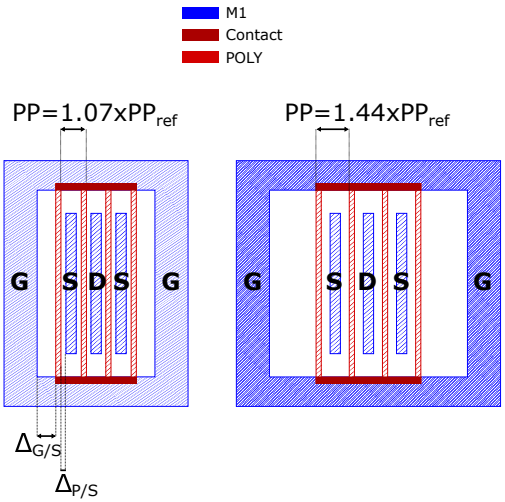


Figure 4.4: RF transistor layout in 16FF with PP = 1.07 × PP_{ref} and PP = 1.44 × PP_{ref}.

M9, whereas AL is not used because of its larger resitivity. The question arises as to whether it is better to have 1) less NDP unit devices with more fingers or 2) more NDP unit devices with less fingers. If the horizontal width of the unit device scaled proportionally to N_{fing} one could easily say that approach 2) is better, since the interconnections in the lower metal layers are minimized. Unfortunately this is not the case because there is an additional spacing between the unit cells caused by the substrate contact around the transistor. Therefore if a small N_{fing} is chosen, the transmission-line effect described in section 4.1 comes into play at relatively low M. The results in Figure 4.5 show that the extreme N_{fing} values of 6 and 32 allowed by the PCell are sub-optimal, whereas for N_{fing} between 12 and 20, fairly good peak FoM_{3dB} is achieved. Based on the available data, N_{fing} = 16 and M = 8 are chosen. In order to avoid an unnecessary increase of the horizontal size, a single substrate contact was drawn manually all around the entire NDP array instead of using the standard PCell option with an individual contact around each transistor.
In conclusion, the unit cell in 16FF utilizes RF transistors with N_{fins} = 12, N_{fing} = 16, $L_g = 1.5 \times L_{g,ref}$, $PP = 1.44 \times PP_{ref}$, corresponding to an active width $W_u \sim 19.2\,\mu m$, where "u" stands for "unit device". The correct way to compute W_u in a FinFET technology is explained in detail in section 2.2.4.

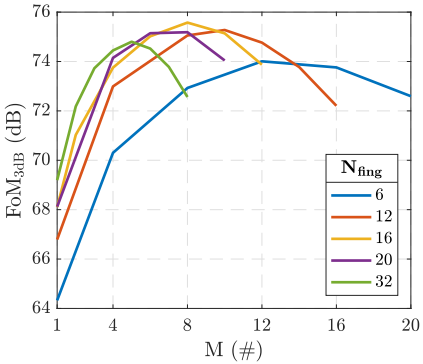


Figure 4.5: FoM_{3dB} vs M of PA core in 16FF at 80 GHz with $N_{fins} = 12$, $L_g = 1.5 \times L_{g,ref}$, $PP = 1.44 \times PP_{ref}$ for different values of N_{fing} based on post-layout simulations.

Based on this analysis the PA core consists of M = 8 unit devices, resulting in a total core width $W_{tot} = 153.6\,\mu m$. This value could in principle change once the impact of the output MNW is considered, as discussed in Chapter 5. A three-dimensional view of the PA core with M = 8 is shown in Figure 4.6. With $V_{gs} = 0.5\,V$, a bias current density of $570\,\mu A/\mu m$ is obtained.

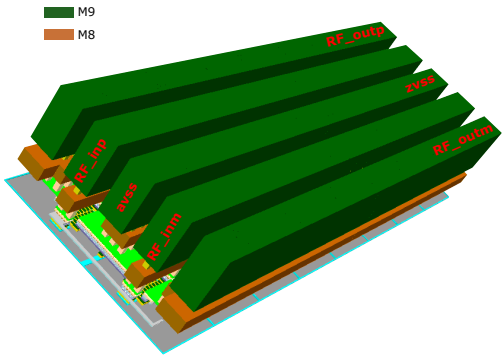


Figure 4.6: Three-dimensional view of the PA core with M = 8 in 16FF.

The FoM_{3dB} loadpull circles for the designed core displayed in Figure 4.7a show an elliptical shape similar to that of the sample curves from Figure 2.4. Furthermore Figure 4.7b shows that the location of the $Z_{L,opt}$ value which

maximizes FoM_{3dB} is a trade-off among those which maximize $P_{out,3dB}$, G_{3dB} and PAE_{3dB} respectively.

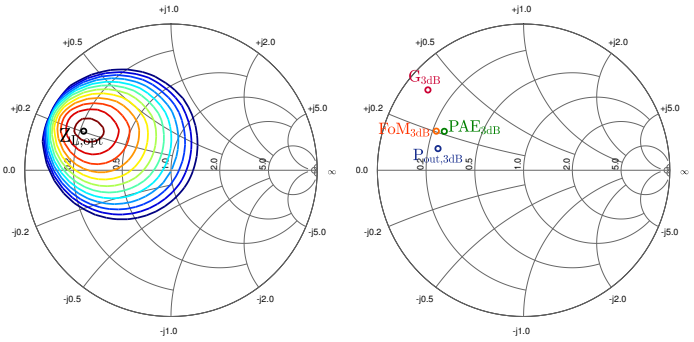


Figure 4.7: (a) Loadpull circles of FoM_{3dB} and (b) comparison of the optimum load impedances for G_{3dB}, $P_{out,3dB}$, PAE_{3dB} and FoM_{3dB} on a Smith-Chart for the optimized PA output stage in 16FF.

4.3.2 Design in 22SOI

The design in 22SOI follows a logic similar to the one used for 16FF but is slightly more complex due to the larger number of design parameters involved. The first step is to determine the optimum back-bias voltage V_{bb} (see section 2.2.3), which in the utilized device can take on values between 0 V and 2 V. Figure 4.8a shows that as V_{bb} increases, the output power for a given value of V_{gs} also increases due to the lower V_t. Figure 4.8b shows that the effect on FoM_{3dB} is merely a shift of the peak of to a lower V_{gs}, without any significant improvement in the value itself. For this reason it is decided to use $V_{bb} = 0\,V$. The optimization of the transistor parameters starts once again from those whose influence can be accurately described by pre-layout simulations, first and foremost the gate pitch PP. The results in Table 4.3 refer to a unit NDP with the same equivalent geometrical parameters and active width ($W_u = 19.2\,\mu m$) as the 16FF device determined in section 4.3.1. It is observed that, differently from 16FF, in 22SOI a larger PP not only does not lead to improved FoM_{3dB}, but causes even a slight degradation. The device with $PP = 2 \times PP_{ref}$ achieves indeed a significant reduction of R_s and R_d, but shows two fundamental issues.

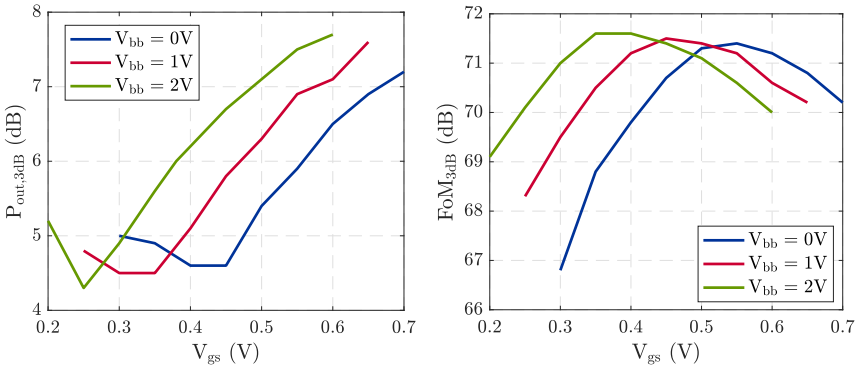


Figure 4.8: (a) $P_{out,3dB}$ and (b) FoM_{3dB} vs V_{gs} of 22SOI PA core with $W_f = 1.2\,\mu m$, $N_{vf} = 1$, $N_{fing} = 16$, $L_g = 1.33 \times L_{g,ref}$, $M = 1$ (pre-layout) at 80 GHz for different values of V_{bb}.

The first is that the spacing $\Delta_{P/S}$ between the gate fingers and the S/D contacts does not increase with respect to $PP = 1 \times PP_{ref}$ (see Figure 4.9), so that the parasitic capacitances C_{gs}, C_{gd} and C_{ds} do not decrease. The second is that the distance $\Delta_{G/S}$ between the G and S/D contacts increases significantly, resulting in substantial overhead on R_g. Adding a gate pitch option at half-way between $1 \times PP_{ref}$ and $2 \times PP_{ref}$ following the same guidelines used in 16FF is expected to improve the performance. Since such a custom layout would require to build the device model from scratch, this is left as a possible subject of future work. The optimum V_{gs} for $PP = 1 \times PP_{ref}$ is 0.53 V, which corresponds once again to class A operation, considering that the device has $V_t \sim 0.23$ V. In this analysis $V_{gs} = 0.5$ V is taken for simplicity and "compatibility" with 16FF, as it shows only a minor performance difference compared to 0.53 V.

Table 4.3: Performance of PA core in 22SOI with $W_f = 1.2\,\mu m$, $N_{vf} = 1$, $N_{fing} = 16$, $L_g = 1.33 \times L_{g,ref}$, $M = 1$ and different values of PP at $f_0 = 80$ GHz (pre-layout).

PP	**V_{gs} (V)**	**$P_{out,3dB}$ (dBm)**	**G_{3dB} (dB)**	**PAE_{3dB} (%)**	**FoM_{3dB} (dB)**
$1 \times PP_{ref}$	0.53	5.8	11.5	39.7	71.3
$2 \times PP_{ref}$	0.56	6.6	10.5	40.1	71.2

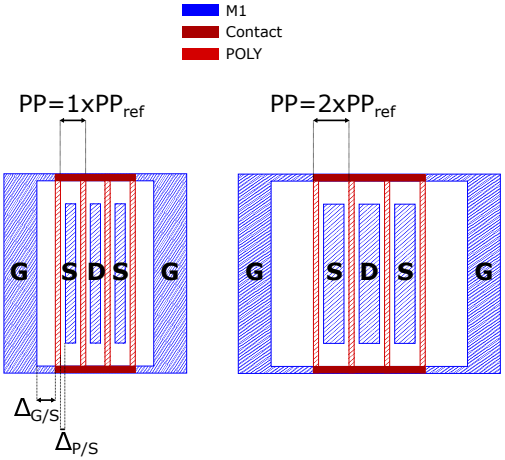


Figure 4.9: RF transistor layout in 22SOI with PP = 1 × PP_{ref} and 2 × PP_{ref}.

Figure 4.10a shows the behavior of FoM_{3dB} as a function of the total finger width $W_{f,tot} = N_{vf} \times W_f$, where $N_{vf} = 1, 2, 3$ and W_f is scaled to keep $W_{f,tot}$ constant. It is observed that choosing $N_{vf} = 2$ or 3 allows the use of a significantly larger $W_{f,tot}$ than it would be possible with $N_{vf} = 1$, as the onset of the performance degradation is shifted to much larger values of $W_{f,tot}$. Even though they are slightly below the peak of FoM_{3dB}, the values $W_{f,tot} = 1.8\,\mu m$ and $N_{vf} = 3$ are chosen to avoid excessive degradation of PAE. As far as L_g goes, Figure 4.10b shows that $L_g = L_{g,ref}$ delivers the highest FoM_{3dB}. This can be explained by the larger bias current ($I_D \propto W/L$), which results in higher $P_{out,3dB}$.
The optimization of the parameters which require a post-layout simulation leads to an important problem tackled in this work, that is whether or not the more expensive metal stack option MO2 brings significant benefits over MO1 (see Figure 2.12b) [RNC+20]. For each metal stack option the input and output terminals of the PA core are routed up to the top available metal layer, i.e. M9 for MO1 and M10 for MO2, to minimize the interconnection resistance and inductance. Figure 4.11 shows that using MO1 the peak FoM_{3dB} is 77.7 dB and occurs for M = 6, whereas using MO2 it is 78.3 dB and occurs for M = 8, which corresponds to a 0.6 dB improvement. The difference however increases with M, and if a large P_{out} is required by the application, the improvement delivered by MO2 becomes substantial. The metal stack has also an impact on

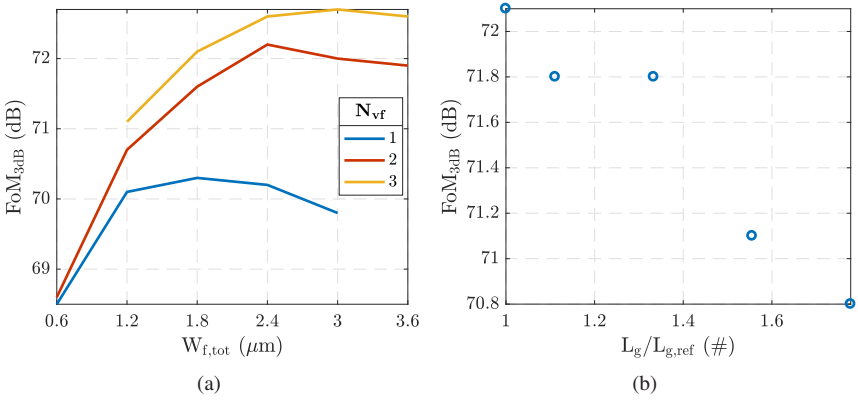


Figure 4.10: FoM_{3dB} of PA core in 22SOI, sweep over (a) W_f, N_{vf} (L_g = 18 nm) and (b) over L_g (W_f = 0.6 µm, N_{vf} = 3), based on pre-layout simulations at 80 GHz with N_{fing} = 16, PP = 1 × PP_{ref}, M = 1, V_{gs} = 0.5 V.

the output MNW, as discussed at length in Chapter 5.

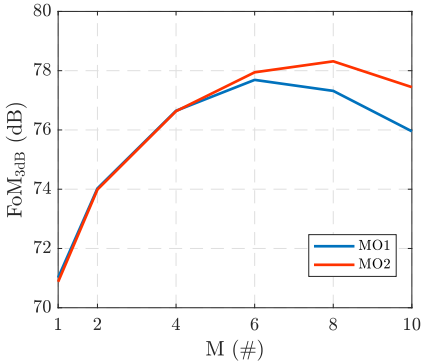


Figure 4.11: Performance of PA core in 22SOI with W_f = 0.6 µm, N_{vf} = 3, N_{fing} = 16, L_g = $L_{g,ref}$, PP = 1× at 80 GHz utilizing MO1 and MO2 (post-layout).

A post-layout analysis utilizing MO2 with joint sweep of N_{fing} and M similar to that shown in Figure 4.6 for 16FF leads once again to the choice of N_{fing} = 16 and M = 8. Differently from 16FF, each transistor has its individual substrate contact because the standard PCell of the component does not offer the option

to leave it out. In order to limit the size of the NDP array, the substrate contacts of adjacent devices are overlappred laterally. In conclusion, the optimum unit device in 22SOI has $W_f = 0.6\,\mu m$, $N_{vf} = 3$, $N_{fing} = 16$, $L_g = 18\,nm$, $PP = 1 \times PP_{ref}$, which corresponds to a unit device active width $W_u = 28.8\,\mu m$. As shown in Figure 4.12, the core has M = 8, subject to change once the effect of the output MNW is considered. With $V_{gs} = 0.5\,V$ the resulting bias current density is $597\,\mu A/\mu m$.

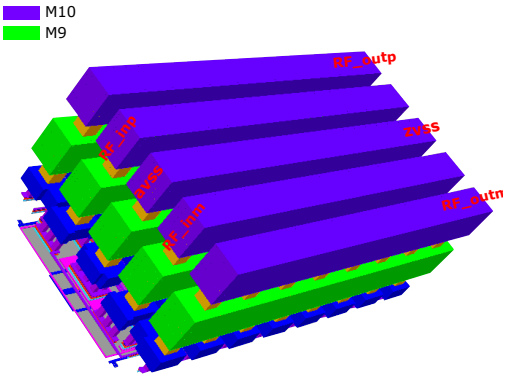


Figure 4.12: 3D view of the PA core with M = 8 in 22SOI with MO2.

4.4 Reliability Considerations

Reliability is a fundamental constraint to take into account in the design of PAs and, in general, of circuits which entail high power levels. While this aspect is sometimes neglected in academic contexts, it should always be considered in an industrial scenario to ensure that the lifetime of the designed system complies with the requirements. One mechanism which can cause the failure of an IC is the voltage stress at the terminals of the active devices, which can lead to the irreversible breakdown of the gate oxide or of the substrate parasitic diodes. Another one is a high current level in the metal interconnects, which can result in Electromigration. The voltage stress issue is caused by the fact that when operating at full power, the voltage swings across the terminals of the transistor are close to the maximum allowed values, defined by the Safe Operating Area (SOA). While V_{gs} and V_{ds} are controlled and known to be in

the ranges $(0V, 2V_{gs})$ and $(0V, 2V_{DD})$ respectively, V_{gd} is simply the result of the swings on the other two terminals and as such is more likely to fall out of the SOA. One of these voltages exceeding the allowed limits can lead to hot-carrier stress, with considerable degradation of the performance, or to gate oxide breakdown, normally resulting in permanent failure of the device [JF13]. From this point of view no major differences are expected between 16FF and 22SOI, since they have the same nominal supply voltage. Transistors in non-SOI technologies feature two substrate pararasitic diodes, one between source and bulk (D_{sb}) and one between drain and bulk (D_{sb}). The voltage swing across these diodes can give rise to junction breakdown. In CS amplifiers like those analyzed in this work D_{sb} does not play any role because bulk and source are shorted, whereas D_{db} is exposed to V_{ds} and could in principle undergo breakdown. In this work only the Electromigration issue is addressed due to its increasing relevance in deeply scaled CMOS nodes. The rest of this section is devoted to the analysis of this phenomenon in the 16FF and 22SOI PA unit devices designed in sections 4.3.1 and 4.3.2.

4.4.1 Electromigration: Analysis Techniques

Electromigration (EMG) is a phenomenon consisting in the displacement of atoms within the conducting material of the interconnects caused by high current levels. It can lead to an alteration of the interconnect resistance or, in the worst case, to unwanted open or short circuits. The average time required for a certain metal or via interconnect to fail as an effect of EMG when it is exposed to a current density J at a temperature T is given by the mean time to failure MTTF (J, T). This can be calculated using Black's equations [Bla69]:

$$\mathrm{MTTF}\,(J, T) = \frac{A}{J^n} \exp\left(\frac{E_a}{k_B T}\right) \tag{4.2}$$

where E_a is the activation energy, k_B is Boltzmann's constant, A is a constant which depends on the properties of the material and the geometry of the interconnects and n is a technology-dependent scaling exponent. The statistical behavior of the wear-out process over time is typically described using a log-normal distribution with density function:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(t) - \ln(\mathrm{MTTF}(J,T)))^2}{2\sigma^2}\right) \tag{4.3}$$

where t is time and σ is the standard deviation. Integrating (4.3) the cumulative failure rate f_r at time t_0 is obtained:

$$f_r(t_0, J, T) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\ln(t_0) - \ln(\mathrm{MTTF}(J,T))}{\sqrt{2}\sigma}\right)\right) \tag{4.4}$$

There exist commercial sign-off tools which utilize this statistical model to determine potential EMG issues in a given circuit layout. Taking as constraint a lifetime t_{life} and the maximum failure rate $f_{r,max}$, the tool computes the maximum allowed current density J_{lim} for each metal or via layer using (4.4) with $t_0 = t_{life}$ and $f_r = f_{r,max}$. The actual current density is simulated using an extracted RC netlist of the layout and this value is compared to J_{lim}, thus detecting a pass or a fail. In this work $t_{life} = 5$ yr and $f_r = 1$ dpm (defect per million) are assumed, which means that a maximum failure rate of one device out of one million is tolerated when the circuit undergoes the specified stimulus continuously for 5 years.
Due to its strong impact on J_{lim}, the temperature T of the interconnect should be estimated as accurately as possible. It can be expressed as:

$$T = T_{env} + \Delta T_{JH} + \Delta T_{SH} \tag{4.5}$$

where T_{env} is the environment temperature, ΔT_{JH} the temperature increase caused by the Joule effect and ΔT_{SH} the temperature increase caused by the transistor Self Heating (SH) effect. T_{env} is application-specific and should be the highest value for which circuit operation must be guaranteed. Since in this work no application is specified, $T_{env} = 80\,°C$ is considered, which is a typical value for many applications. The Joule effect is the temperature increase caused by the current flow in the conductor, which depends on the Root-Mean-Square (RMS) value of the current density J_{rms}. The SH effect is the heating

of the active devices and resistors, which cause a temperature increase in the overlaying metal interconnections given by:

$$\Delta T_{SH} = C_{sh,i} \Delta T_{feol} \tag{4.6}$$

where ΔT_{feol} is the temperature increase of the transistor or resistor in the Front End of Line (FEOL) and $C_{sh,i}$ is the SH coefficient of the i-th metal layer or via. In this work one of the above mentioned commercial tools for reliability sign-off is utilized for the analysis of the NDP unit devices in 16FF and 22SOI designed in sections 4.3.1 and 4.3.2 respectively. The simulation setup consists of the NDP with input and output matched to 100 Ω differential source and load by means of ideal Inductive-Capacitive (LC) MNWs centered at 80 GHz, as shown in Figure 4.13. Running a transient analysis under a given bias current I_d and an input RF signal at 80 GHz with power P_{in} one can determine the current density at every point of the layout and compare it with J_{lim}.

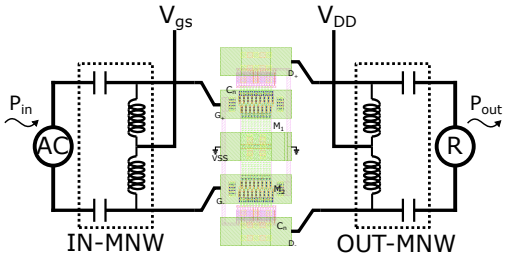


Figure 4.13: Testbench for EMG analysis of the NDP unit device.

In order to cover all the operating scenarios of the PA, two distinct analyses are performed. The first one with DC excitation only, sweeping the gate bias voltage V_{gs} to obtain different bias currents I_d, including the value $I_{d,bias}$ for which the stage has been designed. The second analysis is run with fixed DC bias $I_d = I_{d,bias}$ and an RF signal at 80 GHz with variable input power P_{in}. A typical outcome of the DC analysis is shown in Figure 4.14, where the numbers on the vertical axis correspond to the various metals and vias. For each value of I_d the presence of the horizontal blue line denotes an EMG violation on the corresponding layer, whereas the vertical black line indicates $I_d = I_{d,bias}$. A set of horizontal lines completely located on the right-hand side of the black vertical

line denotes a full pass, whereas lines breaking into the left-hand side indicate a potential concern on that specific metal or via. The same considerations hold for the Alternate Current (AC) analysis, whose outcome is a similar plot but as a function of P_{out}, with the vertical black line corresponding to $P_{out,3dB}$.

4.4.2 Electromigration in 16FF and 22SOI

Significant differences between planar and FinFET technologies have been observed in the way the effects described in section 4.4.1 come into play. As explained in section 2.2.4, in FinFET technologies the total effective width $W_{tot,eff}$ of an active device is typically larger than the drawn width $W_{tot,d}$, in the case of 16FF by a factor of 2. The main consequence is that, assuming the same $W_{tot,eff}$ for the two types of process, in FinFET the current density in the interconnects is larger than in planar. Another key difference is that in FinFET the heat generated in the fins can not be sufficiently dissipated into the substrate due to the limited contact surface [KBL14], as clearly visible comparing Figures 2.11a and 2.11b. Based on these considerations, EMG is expected to be more of concern in 16FF than in 22SOI.
The analysis technique described in section 4.4.1 was applied to the NDP unit devices in 16FF and 22SOI designed in section 4.3. The scope is not only to check whether these are clean from the point of view of EMG, but also to justify some layout choices that have been made and provide a comparison between the two technologies. Figures 4.14a and 4.14b show the results with only DC excitation at $T_{env} = 80\,°C$ for an NDP in 16FF based on active devices which utilize the standard and custom M1-M3 routing respectively (see section 4.3.1). The layout with standard routing shows violations on via1, m2, m3 and via3 and marginal violations on via2, m4, via4 and via5 for $I_d = I_{d,bias}$. Using the transistor layout with custom routing significant improvement is achieved but the violations are not completely resolved. A moderate relaxation of t_{life} and f_r is expected to clear the residual violations, but unfortunately the model files in 16FF do not allow to specify values other than $t_{life} = 5\,yr$ and $f_r = 1\,dpm$.
A similar analysis is performed in 22SOI to compare NDPs with $PP = 1 \times PP_{ref}$ and $PP = 2 \times PP_{ref}$. Increasing the gate pitch helps to relax the EMG limits by allowing the current in the interconnects above the active device to spread over a larger area. This is confirmed by the plots in Figures 4.15a and 4.15b, which show that for $I_d = I_{d,bias}$ using $PP = 2 \times PP_{ref}$ instead of $PP = 1 \times PP_{ref}$ solves

most of the violations, with some residual issues only on m4.
A comparison of Figures 4.14b and 4.15b shows that a similar number of fails are reported in the two technologies in spite of the larger W_u of the 22SOI device. The expectation that 22SOI be much better than 16FF in terms of EMG however is not verified. The reason is that the behavior of J_{lim} as a function of T in 22SOI follows exactly Black's law (4.2), whereas in 16FF it shows significant relaxation at high T. Moreover the $C_{sh,i}$ coefficients are larger in 22SOI than in 16FF so that, even though ΔT_{feol} is larger in 16FF, ΔT_{SH} is larger in 22SOI. These differences in the modelling of EMG are due to the fact that the 16FF and 22SOI processes come from different foundries. Since each foundry can decide to adopt a more or less conservative approach, a direct comparison is difficult.

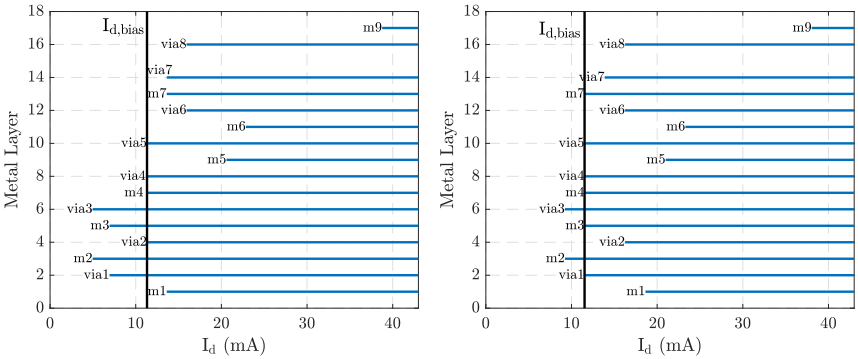


Figure 4.14: EMG plots as a function of I_d with $P_{in} = 0\,W$ at $T_{env} = 80\,°C$ for unit NDPs in 16FF using active devices with M1-M3 routing (a) from foundry and (b) custom.

The AC analysis is conducted for both technologies setting $I_d = I_{d,bias}$ and sweeping P_{in}. As shown in Figure 4.16a, in 16FF there are some fails at low P_{in} which are resoved at larger P_{in}. This might appear counter-intuitive at first, but is in fact expected. Indeed the output voltage and current waveforms of a single transistor terminated on an impedance $Z_L = (V_0/I_0)e^{j\phi}$ are given by:

$$V_{out}(t) = V_{dd} + V_0 cos(\omega_0 t + \phi) \tag{4.7}$$

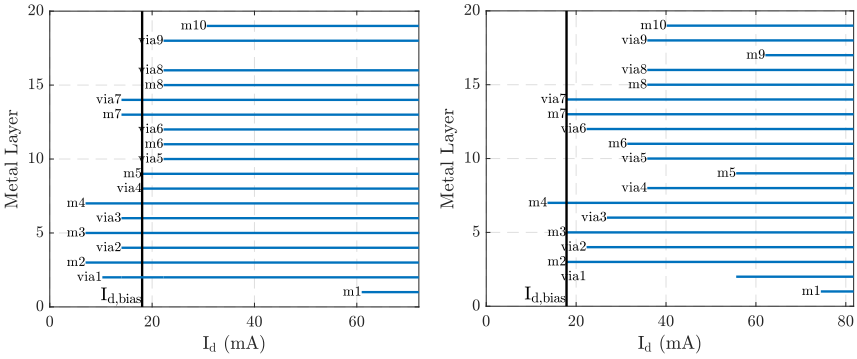


Figure 4.15: EMG plots as a function of I_d with $P_{in} = 0\,W$ at $T_{env} = 80\,°C$ for unit NDPs in 22SOI using active devices with (a) PP = $1 \times PP_{ref}$ and (b) PP = $2 \times PP_{ref}$.

$$I_{out}(t) = I_d - I_0 \cos(\omega_0 t) \tag{4.8}$$

The DC and AC terms of the current waveform have opposite signs because the former flows into the device and the second into the load. The average power dissipated in the transistor reads:

$$P_{d,tr} = \frac{1}{T}\int_0^T V_{out}(t) \cdot I_{out}(t)\,dt = V_{dd} I_d - \frac{V_{rms}^2}{|Z_L|}\cos(\phi) \tag{4.9}$$

where $V_{rms} = V_0/\sqrt{2}$. Since $\cos(\phi)$ is positive for any passive load, $P_{d,tr}$ is maximum with $P_{in} = 0\,W$ and decreases with increasing P_{in}, which results in less severe ΔT_{SH}. This makes total sense physically, as the RF power is generated in the transistor and dissipated on the load. At very large P_{in} however ΔT_{JH} tends to become dominant, which might result into new fails, even on layers that had been resolved at large P_{in}. This effect is clearly observed on via3 in 16FF, which fails at low power, gets resolved for P_{in} around 4 dBm and fails again above 6 dBm or so (Figure 4.16a). In conclusion the DC and AC

analyses are both necessary and the observed violations should be combined. In 22SOI none of the low-power fails gets resolved at high power, as shown in Figure 4.16b, which can be explained once again by the more stringent limits compared to 16FF.
In 16FF the active device with custom routing was adopted in the design because it shows better performance as well as better robustness to EMG compared to the one with standard routing (see section 4.3.1). In 22SOI the device with $PP = 2 \times PP_{ref}$ shows better robustness to EMG compared to the one with $PP = 1 \times PP_{ref}$ but worse performance, especially for large values of M, as discussed in section 4.3.2. For this reason, and partly because the EMG analysis was performed only after the design and tape-out were completed, the layout with $PP = 1 \times PP_{ref}$ was used in the final design.

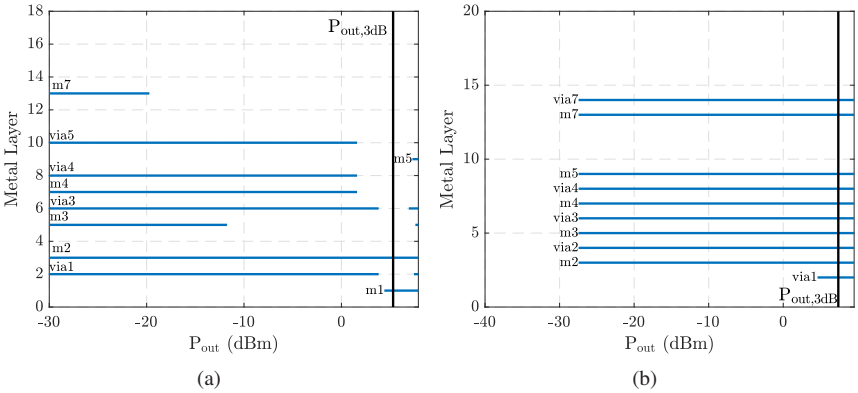


Figure 4.16: EMG plots as a function of P_{in} for $V_{gs} = 0.5\,V$ ($I_d = I_{d,bias}$) at $T_{env} = 80\,°C$ for unit NDPs (a) in 16FF with custom M1-M3 routing and (b) in 22SOI with $PP = 1 \times PP_{ref}$.

4.5 Summary

In this chapter the design of the amplifying stages of the PA was discussed in detail. The ITRS figure of merit evaluated at the 3dB-compression point was identified as the proper metric to drive an algorithmic optimization methodology and applied to the design of NDP-based output stages with optimum performance in the 16FF and 22SOI processes. The design procedure has high-

lighted the impact of the transistor layout features on the circuit performance, above all the gate finger width, the number of gate contacts, the gate pitch, the routing of the bottommost metal layers and the number of unit device cells in the stage. In 22SOI it was found that the metal stack option with one additional ultra-thick metal brings a modest improvement of FoM_{3dB} of about 0.6 dB and only above a certain value of the multiplicity. Furthermore the electromigration effect and its different impact on planar and FinFET technologies has been discussed with the support of simulations of the unit NDPs in 16FF and 22SOI. Some mitigation measures have been proposed for the two technologies, highlighting the fact that they might lead to performance degradation in some cases. Due to the slightly different methodologies used to model the wear-out in the two processes, the outcome of this analysis can not be regarded as a comparison of the physical limitations of the two technologies. Nonetheless, it provides limiting values which can be utilized for the circuit design.

5 Design of the Matching Networks

Impedance MNWs are critical components in analog circuits, especially at RF and mmW frequencies, as they ensure maximum power transfer through the circuit and prevent reflection phenomena. As any other passive circuits they introduce power loss, which proves particularly harmful in the case of the output MNW, since it causes further reduction of G_2 in (2.6). It is well-known that the Insertion Loss (IL) of the MNW increases with the Impedance Transformation Ratio (ITR), a phenomenon which is referred to as "fundamental limitation of impedance matching networks" throughout this work. While this principle is undoubtedly true, an intuitive explanation and a comprehensive statement on its validity limits are missing in the published literature. This is what it is attempted to provide in section 5.2, pointing out the main differences between two of the most common types of MNW topologies, namely the LC and transformer-based networks. Another critical problem tackled in this Chapter is that of predicting the IL caused by the MNW in the PA without simulating the entire circuit. In section 5.3 it is shown that the formulas proposed in the available research work typically correspond to one of the microwave power gains, whose definitions are briefly recalled in section 5.1, and hold only under certain assumptions on the termination impedances. As a conclusion, a large-signal generalization of the transducer power gain is identified as the most suitable figure of merit to achieve global efficiency optimization, even under non-linear operation of the active stages. In section 5.4 a layout optimization methodology for transformer-based MNWs is outlined and applied to the synthesis of output MNWs for the PA cores in 16FF and 22SOI designed in Chapter 4. This analysis sheds some light on many interesting design aspects, such as the advantages of using transformers with 1:2 turn ratio and operating above the Self-Resonant Frequency (SRF). Furthermore, it allows to extend the comparison of the two available metal stack options in 22SOI started in Chapter 4 to include the impact on the passive components. In section 5.5 a study on the standalone characterization of transformers is carried out and validated by means of suitable test structures in the 16FF process. Finally in section 5.6 the conclusions of this Chapter

are summarized. The main results presented in sections 5.1 through 5.3 are extracted from [5], those in section 5.4 related to 16FF from [3] and those in section 5.5 from [2].

5.1 The Microwave Power Gains

Any two-port electrical network can be represented as a black box driven by a source with impedance Z_s and terminated a load with impedance Z_L, as shown in Figure 5.1. The black box can be any active or passive electrical network and can be described by its S-parameters under linear operating conditions. In the case under analysis the black box is an impedance MNW, Z_s is the complex output impedance of the preceding PA stage and Z_L is the complex input impedance of the following stage. In the case of the output MNW, Z_L is typically the 50 Ω antenna impedance.

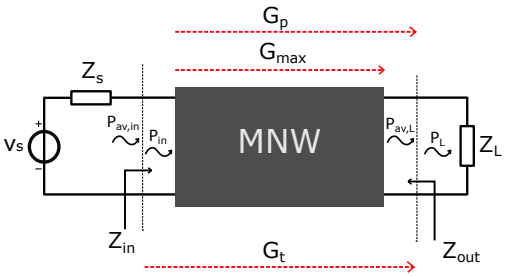


Figure 5.1: Schematic of a MNW with source and load impedances Z_s and Z_L (reprinted from [5] ©2021 IEEE).

It is possible to define three types of power gains, namely the available gain $G_a = P_{av,L}/P_{av,in}$, the operating gain $G_p = P_L/P_{in}$ and the transducer gain $G_t = P_L/P_{av,in}$ [GP17]. P_{in} is the power delivered to the input, $P_{av,in}$ the available power at the input, P_L the power delivered to the load and $P_{av,L}$ the available power at the load. Another fundamental definition is G_{max}, namely the common value of G_a, G_p and G_t obtained when $P_{in} = P_{av,in}$ and $P_L = P_{av,L}$, which corresponds to the maximum power transfer condition. The expressions of the power gains [GP17] are reported below, with the exception of G_a, which is not of interest for our analysis.

$$G_t(\Gamma_s, \Gamma_L) = \frac{|S_{21}|^2 \left(1 - |\Gamma_L|^2\right) \left(1 - |\Gamma_S|^2\right)}{\left|(1 - \Gamma_L S_{22})(1 - \Gamma_S S_{11}) - S_{12} S_{21} \Gamma_S \Gamma_L\right|^2} \tag{5.1}$$

$$G_p(\Gamma_L) = \frac{|S_{21}|^2 (1 - |\Gamma_L|^2)}{|1 - S_{22}\Gamma_L|^2 - |S_{11} - \Delta S \Gamma_L|^2} \tag{5.2}$$

$$G_{max} = \frac{|S_{21}|}{|S_{12}|}(K_f - \sqrt{K_f^2 - 1}) \tag{5.3}$$

$$K_f = \frac{1 + |\Delta S|^2 - |S_{11}|^2 - |S_{22}|^2}{2|S_{21}||S_{12}|} \tag{5.4}$$

In these equations S_{ij} are the 2-port S-parameters, $\Delta S = S_{11}S_{22} - S_{12}S_{21}$ is the determinant of the S-matrix, Γ_s and Γ_L are the source and load reflection coefficients and K_f is the stability factor. The reflection coefficient Γ_i associated to the impedance Z_i is defined as:

$$\Gamma_i = \frac{Z_i - Z_0}{Z_i + Z_0} \tag{5.5}$$

where Z_0 is the normalization impedance of the S-matrix.
While the power gains were originally defined to quantify the gain of active circuits, in this case they are used to estimate the loss of a passive network. Based on the definitions above, they have the following straightforward physical interpretation:

- G_{max} is the power gain obtained when the input and output of the MNW are conjugately matched to the source and load respectively.
- G_p is the power gain obtained when the input is conjugately matched to the source but with generic matching condition at the load, hence taking mismatch loss at the output into account. If the circuit is unconditionally stable, there exists only one value of Γ_L for which $G_p = G_{max}$.
- G_t is the power gain obtained with generic matching conditions at the source and at the load, hence taking mismatch losses at both input and

output into account. If the circuit in unconditionally stable, there exists only one set of values (Γ_s, Γ_L) for which $G_t = G_{max}$.

In order to quantify the insertion loss of a MNW, the power gain corresponding to the actual source and load matching condition of the network should be used [HTC08].

5.2 The fundamental Limitation of Impedance Matching Networks

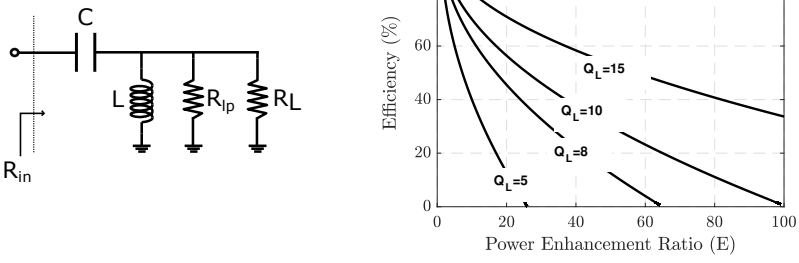
As mentioned in the Introduction to this Chapter, the fundamental limitation of the impedance MNWs consists in the fact that the IL increases with the ITR. This phenomenon is typically observed in deeply scaled CMOS technologies, since the progressive reduction of the supply voltage V_{DD} in more advanced nodes results in a similar decrease of the maximum voltage swing at the drain of the active devices. In order to keep the output power at a fixed desired level, the current swing has to increase. This is achieved by increasing the active width of the output stage, which leads invariably to a smaller optimal load impedance $Z_{L,opt}$. Since $Z_{L,opt}$ is typically much smaller than the load presented to the PA, for instance the 50 Ω antenna, a MNW with a large ITR is required, which suffers from low efficiency. With increasing operating frequencies and aggressive technology scaling, this fundamental limitation of impedance MNW is coming back into focus [CS17]. This issue is frequently mentioned in several publications about mmW PAs [SPD+16, HCM16] to justify the use of device stacking and power combining. Most of these works refer to a single but very relevant study by Aoki et al. [AKRH02], which analyzes how LC and transformer-based MNWs are affected by this limitation. Since a full understanding of this topic is critical to determine which types of MNWs require efficiency optimization, in this section the main statements of [AKRH02] are reviewed and discussed. Using the LC MNW as case study, an attempt is made to provide a more intuitive view of the problem. As for the transformer, the applicability of the statements made in [AKRH02] is discussed, comparing the main results with those from later studies.

5.2.1 LC Matching Networks

In [AKRH02] the efficiency is defined as $\eta = P_L/P_{in}$, which corresponds to G_p from section 5.1. Based on this formulation, the efficiency η_{LC} of the single-section LC MNW with series-C/shunt-L topology in Figure 5.2a is derived:

$$\eta_{LC} = \frac{Q_L^2 + 1}{Q_L^2 + \frac{1}{2}\left(r + \sqrt{r^2 + 4Q_L^2(r-1)}\right)} \tag{5.6}$$

The capacitor is assumed to be ideal ($Q_C = \infty$), whereas the inductor is assigned a finite quality factor $Q_L = R_{lp}/(\omega L_p)$, where R_{lp} is the parallel parasitic resistance of the inductor itself. Additional assumptions are that the load R_L is purely resistive and that L and C are chosen to obtain a purely resistive input impedance R_{in}, so that the ITR can be defined as $r = R_L/R_{in}$. A parameter called power enhancement ratio $E = r\eta$ is also introduced, which corresponds to the power enhancement that the MNW would provide if the input and output voltages of the two-port network were referred to an identical impedance level, that is $E = V_{out}^2/V_{in}^2$. In Figure 5.2b it is shown that the efficiency of the MNW degrades with increasing E or, equivalently, with increasing ITR, and improves with increasing Q_L.



(a) (b)

Figure 5.2: Single-section LC MNW: (a) Circuit schematic and (b) plot of Efficiency vs E for different values of Q_L (reprinted from [5] ©2021 IEEE).

While it is obvious that including the parasitic resistance R_{lp} in the model results in efficiency degradation, it might not be so clear why this degradation increases with the ITR. In order to scale down the resistance, the load has to be shunted by a reactive impedance, in this case inductive, to move in the desired direction on the Smith Chart. The more the resistance has to be scaled down, the smaller the shunt inductance has to be. Since a constant Q_L is assumed for the inductor, a smaller L comes with a smaller R_{lp}. This causes the voltage and current division to become increasingly unfavorable for the load, which results in larger power loss. The very same considerations presented so far apply to Transmission Line (TL)-based MNWs, which are nothing else than the distributed counter-part of the LC networks.

5.2.2 Transformers

Integrated transformers consist of two magnetically coupled inductors called "primary" and "secondary" which are typically built on the topmost ultra-thick layers of the metal stack. They are classified based on the geometrical shape of the inductors and on the topology, which can be stacked or interleaved [CNP+19], corresponding to vertical and lateral magnetic coupling respectively. As an example Figure 5.3 shows the three-dimensional view of a stacked octagonal transformer built on metal layers M_{top} and M_{top-1}. Transformers have found very large application in the last couple of decades in differential mmW PAs because they allow to perform impedance transformation, single-ended to differential conversion and DC biasing of the active stage concurrently [BY19, Bev20, TPAG13].
A simple yet accurate lumped model of a transformer [LSC+20] consists of the primary and secondary inductors of inductance L_p and L_s respectively and magnetic coupling factor k_m, as shown in Figure 5.4a. The loss is modeled by means of the series resistances R_p and R_s, which allow to define the Q-factors as $Q_p = (\omega L_p)/R_p$ and $Q_s = (\omega L_s)/R_s$. The stray parasitic capacitances of the primary and secondary coils are modeled by C_p and C_s, whereas C_c models the capacitive coupling between the two coils. The analysis conducted in [AKRH02] is based on a simplified version of this model, shown in Figure 5.4b, which neglects the parasitic capacitances, thus limiting the validity of the results to frequencies significantly smaller than the SRF of the transformer. Although this assumption generally does not hold in the MNWs designed in this

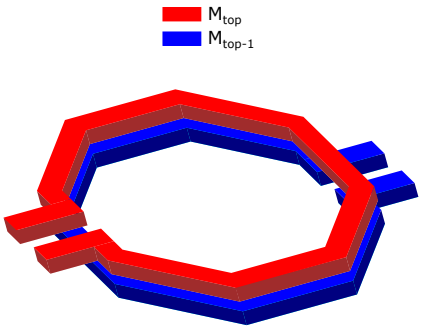


Figure 5.3: Three-dimensional view of a vertically-coupled octagonal transformer.

work, as will be discussed in section 5.4, the model of Figure 5.4b is nonetheless useful to explain how the fundamental limitation applies to transformers.

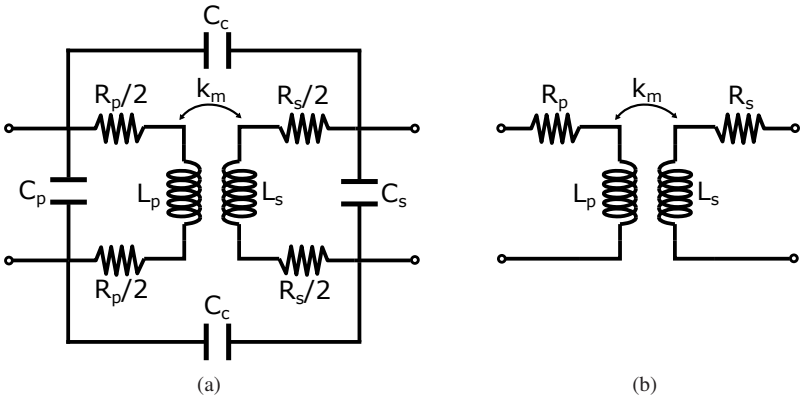


Figure 5.4: Lumped model of a transformer-based MNW at (a) high-frequency and (b) low-frequency (<SRF).

In [AKRH02] it is shown that under the conditions on L_p and on the load impedance Z_L reported in A.1, the maximum transformer efficiency $\eta_{tr,max}$ is obtained:

$$\eta_{tr,max} = \frac{1}{1 + 2\sqrt{\left(1 + \frac{1}{Q_p Q_s k_m^2}\right)\frac{1}{Q_p Q_s k_m^2}} + \frac{2}{Q_p Q_s k_m^2}} \tag{5.7}$$

This expression is independent of the ITR as well as of the individual values of R_{in} and R_L and can be proved to be equivalent to the G_{max} of the transformer. Hence the above-mentioned conditions on L_p and Z_L are those which guarantee simultaneous conjugate match at input and output. While this proves that this condition is theoretically achievable in transformers as opposed to LC networks, it should be also mentioned that the required condition on L_p and Z_L are normally not at all easy to meet. An efficiency estimation comparable to (5.6) can be obtained from the model of Figure 5.4b considering a purely resistive load R_L and without imposing any constraints on the parameter values [TP19]:

$$\eta_{tr} = \frac{P_L}{P_{IN}} = \frac{R_L}{R_L + R_S + R_P \frac{(R_L + R_S)^2 + (\omega L_S)^2}{(\omega k_m)^2 L_p L_s}} \tag{5.8}$$

The input impedance of the transformer reads:

$$Z_{in} = R_{in} + jX_{in} = \left(R_p + \frac{\omega^2 k_m^2 L_p L_s (R_s + R_L)}{(R_s + R_L)^2 + (\omega L_s)^2}\right) + j\left(\omega L_p - \frac{\omega^3 k_m^2 L_p L_s^2}{(R_s + R_L)^2 + (\omega L_s)^2}\right) \tag{5.9}$$

Combining (5.8) and (5.9) an expression of η_{tr} as a function of the input resistance R_{in} is obtained, which can be seen as the counter-part of (5.6) for transformers:

$$\eta_{tr}(R_{in}) = \frac{1}{\left(1 + \frac{R_s}{R_L}\right)\left(1 + \frac{1}{\frac{R_{in}}{R_p} - 1}\right)} \tag{5.10}$$

This expression shows that indeed the efficiency of the transformer does not depend on the ITR [CHW+16] but it does depend individually on R_{in} and

R_L. Quite understandably, the transformer has high efficiency only if $R_s \ll R_L$ and $R_p \ll R_{in}$, where R_{in} and R_L are imposed by the preceding and following amplifying stages respectively. A considerable degradation is expected in the presence of an "unfavorable" impedance environment, that is when the values of R_L or R_{in} are small. A small R_{in} is the typical case of the output stage of a mmW PA in a deeply scaled CMOS technology.

5.3 Figure of Merit for the Insertion Loss

Equations (5.6) and (5.10) presented in section 5.2 for LC and transformer-based MNWs are based on a definition of the IL which is equivalent to G_p. As explained in section 5.1, this power gain assumes that the input of the MNW is matched to the source, typically an amplifying stage of the PA. This matching condition, however, is not automatically fulfilled but has to be enforced during the design phase. For this reason, if the figure of merit is used for a design optimization, in which the trade-off among all requirements leads to unintentional mismatch, G_t is a much better choice. The main issue is that in the case of the output stage of a PA the mismatch loss at the interface with the MNW typically shows a non-linear behavior, whereas in G_t a small-signal behavior is implicitely assumed. In the rest of this section this topic is discussed at length and a solution is proposed which partially overcomes the limitations of G_t.

5.3.1 Literature Review

This subsection presents a review of the main methodologies and figures of merit used in the literature for the design of the MNWs, which leads to the solution proposed in this work.
The methodology proposed in [YMYZ15] focuses on the design of transformer-based MNWs, deriving closed-form equations for L_p, L_s and k_m. This is done using the model of Figure 5.4b and enforcing the condition $Z_{in} = Z_{L,opt}$, where $Z_{L,opt}$ is the optimum loadpull impedance of the amplifying stage. Quite interestingly, no constraints are set on the efficiency of the transformer. This is not necessarily an issue if the impedance environment imposed by the design does not result in very low efficiency, but this is not known a priori.

In several works focusing on transformers as standalone components, $\eta_{tr,max}$ (5.7) or the equivalent G_{max} (5.3) are chosen as the quantity to be optimized [CX05, LKBB09, LKB15]. This approach is effective for estimating the maximum efficiency of the transformer, but may not be as useful in circuit design scenarios, as it implicitly assumes that the MNW is conjugately matched both at the input and at the output. This condition can be indeed achieved with a transformer-based MNW [TP19], at least in principle, but it has to be enforced separately as a design constraint. $\eta_{tr,max}$ is also used in some works on transformer-coupled PAs [VR17, Boe10], sometimes under the name of "passive efficiency", but rather as a guideline than as an optimization parameter.
In [CRN09] it is clearly stated that an output MNW should target loadpull match at the input ($Z_{in} = Z_{L,opt}$) and maximum G_p at the same time. It is also mentioned that G_p is the correct figure of merit to optimize, as opposed to G_{max}, since the load is not by default conjugately matched to the output of the MNW. This approach is valid but it has the dowside of using two different optimization parameters.
In [CPH18] a holistic approach is proposed which co-optimizes the active device size and the MNW to achieve the lowest possible loss in large-signal conditions. This allows for instance to reduce the size of the active device if $Z_{L,opt}$ becomes too small for the MNW to be implemented with good efficiency.
Taking the LC network as case study, in [CS17] it is clarified that one should take into account not only the so-called "loss efficiency" η_{loss}, which is equivalent to G_p, but also the matching efficiency η_{match}, which quantifies the power lost due to the mismatch between the active device and the input of the MNW. This is given by:

$$\eta_{match} = \frac{4R_{in}R_s}{|Z_{in} + Z_s|^2} \tag{5.11}$$

where $Z_s = R_s + jX_s$ is the impedance presented by the output terminals of the active stage and $Z_{in} = R_{in} + jX_{in}$ is the input impedance of the MNW. In order to minimize the end-to-end loss of the network, the quantity to be maximized is:

$$\eta_{tot} = \eta_{match}\eta_{loss} \tag{5.12}$$

The main advantage of this formulation is that it quantifies the overall loss of the MNW through a single figure of merit, making it a valuable tool for addressing our design problem.

5.3.2 Proposed Figure of Merit

In order to clarify the relevance of the methodology from [CS17], the single-section LC network with series-L/shunt-C topology of Figure 5.5 is considered. The operating frequency is $f_0 = 80\,\text{GHz}$, the load impedance $Z_L = 50\,\Omega$ and the source impedance $Z_s = (4 - 3j)\,\Omega$, which are reasonable estimations for the output stage of a CMOS PA at these frequencies. Furthermore constant values of $Q_L = 20$ and $Q_C = 15$ are assumed for the quality factors of the inductor and capacitor. The plots in Figure 5.6 are obtained sweeping the values of L and C to generate all possible values of Z_{in} on the Smith-Chart. The color associated with each Z_{in} point represents the corresponding η_{match}, η_{loss} and η_{tot} in dB, displayed in Figure 5.6a, 5.6b and 5.6c respectively. While η_{match} has by definition a maximum for $Z_{in} = Z_s^*$, η_{loss} shows a decreasing behavior as the impedance moves towards the outer rim of the Smith-Chart. As a result, the maximum of η_{tot} is located at a point which is not exactly $Z_{in} = Z_s^*$, but slightly displaced towards Z_L, as shown in Figure 5.6c.

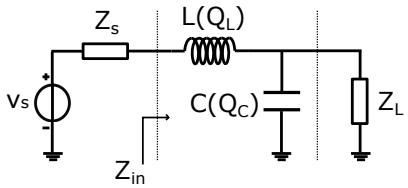
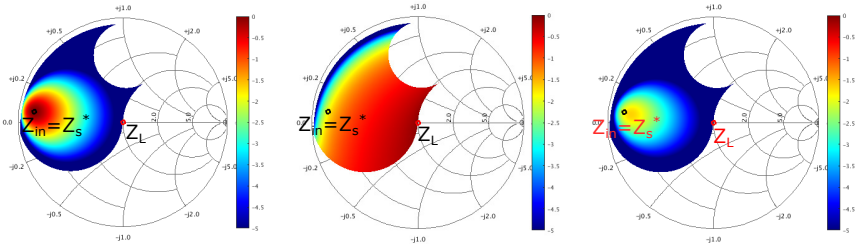


Figure 5.5: LC MNW with series-L/shunt-C topology with source and load impedances (reprinted from [5] ©2021 IEEE).

It can be easily recognized that η_{tot} (5.12) is equivalent to G_t, which is confirmed also in [TP19]. Hence selecting the values of L and C which maximize G_t allows to attain the Z_{in} sweet spot shown in Figure 5.6c. Unfortunately this result holds only as long as the active device operates in small-signal conditions. This is typically not the case for the output stage of a PA, as discussed in detail in section



(a) (b) (c)

Figure 5.6: (a) η_{match}, (b) η_{loss} and (c) η_{tot} in dB as a function of Z_{in} for the LC network of Figure 5.5 at $f_0 = 80\,GHz$ with $Z_L = 50\,\Omega$, $Z_s = (4 - 3j)\,\Omega$, $Q_L = 20$ and $Q_C = 15$ (reprinted from [5] ©2021 IEEE).

2.1.2. In this case η_{match} should be replaced by a large-signal generalization $\eta_{match,LS}$ capable of predicting the elliptical behavior of the loadpull curves:

$$\eta_{tot,LS} = \eta_{match,LS}\,\eta_{loss} \tag{5.13}$$

where η_{loss} is the same as in small-signal conditions. Unfortunately $\eta_{match,LS}$ is normally available only in the form of loadpull data, not as analytic expression. One simple solution is to use (5.11) replacing Z_s with $Z_{s,opt} = Z^*_{L,opt}$, where $Z_{L,opt}$ is the optimum load impedance for the active stage determined by the loadpull simulations:

$$\eta_{match,LS} \sim \frac{4R_{in}R_{s,opt}}{|Z_{in} + Z_{s,opt}|^2} \tag{5.14}$$

This amounts to approximate the elliptical curves of $\eta_{match,LS}$ with circles centered at $Z_{in} = Z_{L,opt}$. If the stage for which the MNW is designed operates in the linear regime, which is normally the case for the driver stages, $\eta_{tot,LS}$ coincides with η_{tot}. One major ambiguity of this definition is that there exist several values of $Z_{L,opt}$, depending on the metric to be optimized, as explained in section 2.1.2. In large-signal conditions $\eta_{tot,LS}$ corresponds exactly to the IL introduced by the MNW only if $Z_{L,opt}$ is the load impedance for maximum G and Z_{in} is exactly equal to $Z_{L,opt}$. Unfortunately none of these two conditions are verified: the former because the utilized design technique selects the load

impedance for maximum FoM_{3dB} and the latter for the reasons discussed above. However it can be shown that the actual IL is monotonic with respect to $\eta_{tot,LS}$, which can therefore be used effectively as a figure of merit to drive the design. In terms of power gains $\eta_{tot,LS}$ is equivalent to a modified version of G_t called $G_{t,LS}$, in which the source impedance Z_s is replaced by $Z_{s,opt}$ (5.15).

5.4 Transformer-Based Output Matching Network Design and Optimization

One of the main advantages of transformer-based MNWs over the LC counterpart is the possibility of performing the impedance transformation and the single-ended to differential conversion concurrently. This aspect does not hold particular relevance for this work because the designed PAs are fully differential, as explained in Chapter 6, but is in general highly beneficial. Since no major differences in terms of IL for a given ITR are anticipated between the two types of networks, it is decided to use transformers due to their higher expected sensitivity to the metal stack profile.

For an effective design methodology, a simple model which relates $G_{t,LS}$ to the electrical parameters of the transformer is required. Over the years a large number of accurate scalable models have been proposed, both of lumped [BSRP06, GJLY06, EGKB07, LKBB12, TPAG13] and distributed [CRN09, HHG+12, NW20, WW21] type. Unfortunately none of these models is sufficiently accurate and yet simple enough to enable an effective design. Moreover, deriving a closed-form expression of $G_{t,LS}$ is not possible even with the highly simplified model of Figure 5.4b. For these reasons in this work a layout optimization approach based on EM simulations [TNM+20] is adopted. It makes use of a PCell, which allows to generate layouts of octagonal stacked transformers with 1:1, 1:2 or 2:1 turn ratios. A rectangular ground ring is also included to take into account the coupling effects to the ground plane and to provide a physical reference for the excitation signals. The adjustable parameters are the coil metal layers M_p and M_s, the ground plane metal layer M_{gp}, the radii r_p and r_s, the trace widths w_p and w_s, the number of turns n_p and n_s, the turn spacings s_p and s_s and the horizontal offset x_{off} between the two inductors (Figure 5.7). The subscripts "p" and "s" refer to the primary and secondary coil respectively. In practice the layouts generated in this section have either 1:1 or 1:2 turn ratio, therefore the only possible metal layer assignment is $M_p = M_{top}$, $M_s = M_{top-1}$

and $M_{gp} = M_{top-2}$, where M_{top} can be the top copper layer or AL (see Figure 2.12). In this way the underpasses of the secondary coil, if present, can be drawn on M_{top-2}.

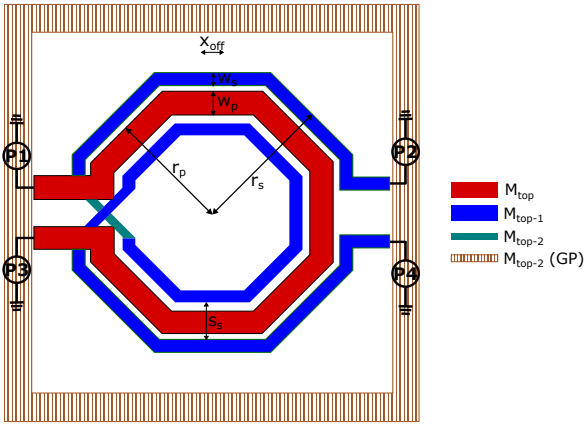


Figure 5.7: Graphical representation of the transformer layout PCell used for the optimization of the MNW. The transformer geometrical parameters are highlighted along with the port configuration used for the S-parameter simulation.

The 4-port S-parameters can be extracted using ports P_1, P_2, P_3 and P_4, which excite the four terminals of the transformer with respect to the ground ring. Starting from (5.1), $G_{t,LS}$ can be expressed as follows:

$$G_{t,LS}(\Gamma_{s,opt}, \Gamma_L) = \frac{\left|S_{dd,21}\right|^2 \left(1 - |\Gamma_L|^2\right) \left(1 - |\Gamma_{s,opt}|^2\right)}{\left|\left(1 - \Gamma_L S_{dd,22}\right) \left(1 - \Gamma_{s,opt} S_{dd,11}\right) - S_{dd,12} S_{dd,21} \Gamma_{s,opt} \Gamma_L\right|^2} \tag{5.15}$$

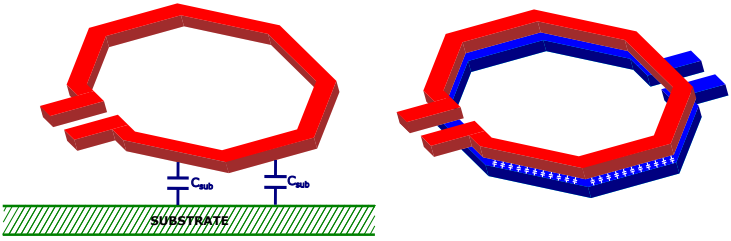
In this expression $S_{dd,ij}$ are the 2-port differential S-parameters of the transformer [FLWL03], which can be extracted from the 4-port single-ended S-parameters [HKGZ10]. The reason why only the differential mode is considered is that the target PA is fully differential, as mentioned at the beginning of this section. For the output MNW the load reflection coefficient Γ_L is the one corresponding to $Z_L = 100\,\Omega$, since each terminal of the secondary of the

transformer is terminated on one 50 Ω port of the VNA.
The optimization methodology is applied to determine the transformer layout with maximum $G_{t,LS}$ for PA cores in 16FF and 22SOI based on the unit device from section 4.3.1. This is done with different values of M, each corresponding to a unique $\Gamma_{L,opt}$. Since M has the strongest impact on $\Gamma_{L,opt}$, the dependency of the IL of the MNW on the impedance environment is also captured and the optimum M can be determined.

5.4.1 Output Transformers in 16FF

In 16FF stacked transformers can be implemented using the metal layer pairs M9/AL or M8/M9. The first option is chosen because it turns out to deliver higher $G_{t,LS}$ thanks to the larger thickness of AL compared to M8. This remains valid in spite of the larger vertical separation between M9 and AL compared to M8 and M9 and of the larger resistivity of aluminum compared to copper. Once the output transformer is added, the FoM_{3dB} vs. M of the PA core undergoes a degradation which follows closely the $G_{t,LS}$ vs. M curve of the transformer (Figure 5.9a), thus validating $G_{t,LS}$ as insertion loss figure of merit.
It is often claimed that transformers should be designed to operate below the SRF [ZHW22], that is the frequency at which the input reactance X_{in} crosses 0 Ω. This is indeed the case for a single inductor, which presents a capacitive impedance above the SRF as an effect of the parasitic capacitive coupling towards the substrate (see Figure 5.8a). This makes it unusable in scenarios in which an actual inductive behavior is required, for instance in an LC MNW. In a transformer the SRF is dictated by the parasitic capacitance between the primary and secondary coils (Figure 5.8b), which is normally much larger than the capacitance between each coil and the substrate. Since X_{in} depends not only on the transformer itself, but also on the termination impedance Z_L of the secondary, an open circuit ($Z_L = \infty$) is assumed in the definition of SRF. Since in a real operating scenario Z_L is always a finite complex value, at frequencies above the SRF the transformer does not necessarily show a capacitive impedance. Even in that case, the impedance transformation action remains effective. Operation above the SRF comes with two issues: (1) the simplified model of Figure 5.4b becomes unsuitable to drive the design and (2) imbalance effects [TP21] kick in due to the significant capacitive coupling between the two coils. The first issue is not of concern because the utilized design method-

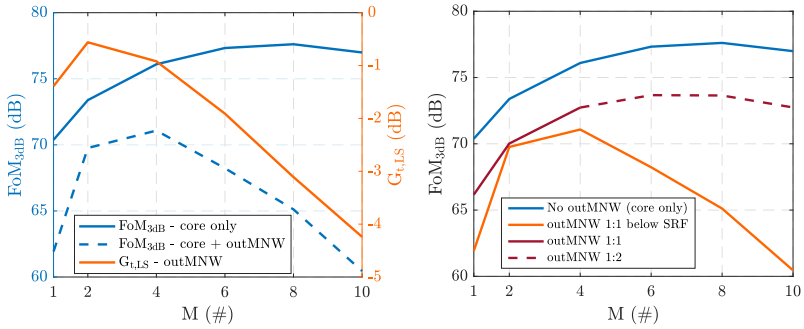
ology does not rely on a circuit model. The second issue is only relevant if the transformer operates in balun mode. This does not apply to the case under analysis, in which a fully differential architecture is used.



(a) (b)

Figure 5.8: Dominant physical mechanisms which determine the SRF: (a) the capacitive coupling to the substrate in a single inductor and (b) the capacitive coupling between primary and secondary in a transformer.

Fig. 5.9a shows that limiting the optimization space to transformers that operate below SRF and with 1:1 turn ratio, huge performance degradation is incurred for large values of M, which limits the usable PA core size to M = 4. Fig. 5.9b shows that removing the SRF constraint and considering also transformers with 1:2 turn ratio results in much better performance. As shown in Table 5.1, the 1:2 turn ratio becomes convenient for $M \geq 4$, as the source resistance R_s becomes low. Transformers with 1:2 turn ratio are typically above SRF at the design frequency due to the increased stray parasitic capacitance associated to the coil with 2 turns. In conclusion, using a transformer with 1:2 turn ratio above SRF, the value M = 8 determined in section 4.3.1 remains the optimum choice even after accounting for the losses of the output MNW.
Plotting the difference ΔFoM_{3dB} between the FoM_{3dB} values with and without MNW against M (Figure 5.10) reveals that the degradation of FoM_{3dB} is minimum in the range between M = 2 and M = 4 and starts increasing monotonically for larger values of M. This behavior is dictated by the source impedance: for small values of M the transformer has to match $Z_L = 100\,\Omega$ to a much larger $Z_{s,opt}$, for large values of M to a much smaller $Z_{s,opt}$, resulting in both cases in strong mismatch. For intermediate values of M instead, good match can be achieved simultaneously at input and output, resulting in low insertion loss.



(a) (b)

Figure 5.9: Impact of output MNW on PA core performance as a function of M in 16FF at 80 GHz (a) using transformer with 1:1 turn ratio below SRF and (b) using transformers above or below SRF, with 1:1 or 1:2 turn ratio.

Table 5.1: Properties of the optimized transformer-based output MNWs from Fig.5.9b (brown curve).

M (#)	$Z_{s,opt}$ (Ω)	Z_L (Ω)	$G_{t,LS}$ (dB)	Turn Ratio	SRF (GHz)
1	91.1-104.7j	100	-0.6	1:1	84
2	50-57.8j	100	-0.5	1:1	93
4	24.7-29j	100	-0.63	1:2	75
6	17.7-16.2j	100	-0.87	1:2	75
8	13.2-11.7j	100	-0.9	1:2	58.5
10	11.2-6.7j	100	-1	1:2	48.7

5.4.2 Output Transformers in 22SOI

As far as 22SOI is concerned, the optimization methodology can be applied to determine whether MO2 brings any benefits over MO1 (see Figure 2.12b) in the IL of the transformer. One obvious advantage of MO2 is that the additional thick copper metal layer allows to build one more coil with high Q. Moreover the upper metal layers M9, M10 and AL are at a larger distance from the lossy substrate, which results in weaker electrical coupling to the substrate itself [NRB02]. The main disadvantage of MO2 is that M9 and M10 show a larger vertical separation compared to M8 and M9 in MO1, which translates into looser magnetic coupling.

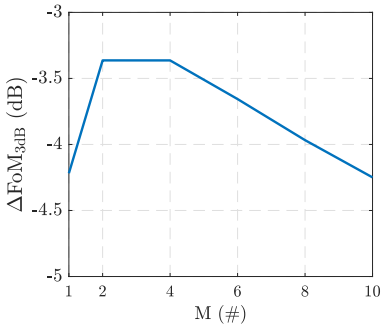


Figure 5.10: ΔFoM_{3dB} vs M at 80 GHz of PA core in 16FF using the transformer-based MNWs of Table 5.1.

It is found that the largest $G_{t,LS}$ is achieved using metal layers AL/M9 with MO1 and M10/M9 with MO2. Figure 5.11a shows that adding the output transformer to the PA core, the peak of FoM_{3dB} shifts from M = 6 down to M = 4 for MO1 and from M = 8 to M = 6 for MO2. This indicates that the limitation caused by the insertion loss of the MNW comes into play slightly earlier than the one caused by the interconnects in the active stage. The reason why this does not happen in 16FF is that W_u is smaller than in 22SOI, so that for a given M the optimum source resistance is higher. The plot of ΔFoM_{3dB} as a function of M in Figure 5.11b reveals that using MO2 the degradation introduced by the transformer is 0.3 dB to 0.6 dB lower than using MO1. This difference is modest and is independent of M. The conclusion is that most of the benefit of MO2 comes from the lower parasitics in the interconnections within the PA core and such benefit becomes noticeable only above a certain value of M. The benefit on the IL of the transformer-based MNW is less relevant and does not depend on the size of the active device. Since MO2 was the standard metal stack option for the tape-out, the simulation and measurement results presented in the rest of this work are based on it.

5.4.3 Technology Comparison

The PA cores with output MNW in the two technologies optimized according to the presented methodology are compared in Figure 5.12. The FoM_{3dB} is plotted

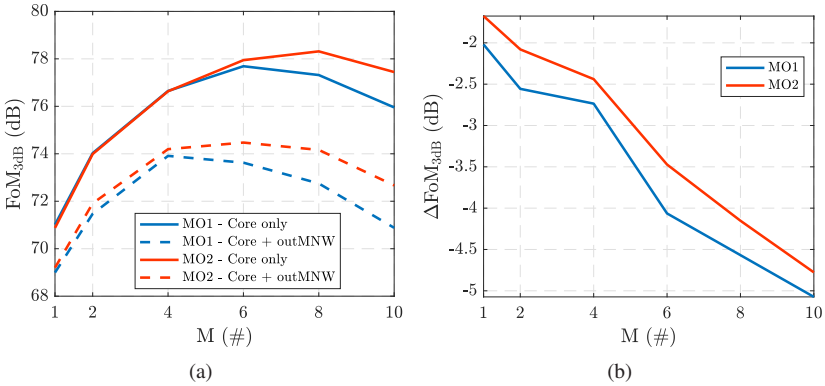


Figure 5.11: Impact of output MNW on PA core performance as a function of M in 22SOI at 80 GHz: (a) FoM_{3dB} vs M and (b) ΔFoM_{3dB} vs M.

against $P_{out,3dB}$ instead of M to allow for a comparison at constant output power. In both processes the output stage with output MNW attains a peak FoM_{3dB} slightly larger than 78 dB. In 22SOI the peak FoM_{3dB} occurs at approximately 1 dB higher $P_{out,3dB}$ thanks to the larger size of the unit device. In 16FF the same FoM_{3dB} is attained at lower $P_{out,3dB}$, but PAE_{3dB} and G_{3dB} are slightly larger than 22SOI, as shown in Table 5.2. This can be explained to a large extent by the optimized the gate pitch, as discussed in sections 4.3.1 and 4.3.2.

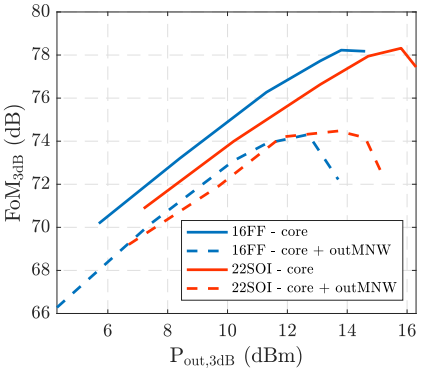


Figure 5.12: Optimized PA core with and without output MNW in 16FF and 22SOI.

Table 5.2: Performance comparison of optimized PA cores in 16FF and 22SOI with transformer-based output MNWs at 80 GHz.

Technology	W_u (µm)	M (#)	W_{tot} (µm)	$P_{out,3dB}$ (dBm)	G_{3dB} (dB)	PAE_{3dB} (%)	FoM_{3dB} (dB)
16FF	19.2	8	153.6	12.7	9.5	25.4	74.3
22SOI	28.8	6	172.8	13.7	9	23.5	74.5

These results lead to the conclusion that, within the limits of the utilized circuit architecture, the two technologies show very similar performance for the design of mmW PAs. The fact that this does not align with the expectations discussed in section 2.2.3 can be explained mainly by three arguments: (1) stacked-FET architectures, which would definitely favor 22SOI, are not used, (2) the RF transistor PCell in 16FF offers better layout options than the one in 22SOI and (3) the possibility of using transformers above SRF makes them less sensitive to the utilized metal stack profile.

5.5 Transformer characterization techniques

In many situations the characterization of a transformer as standalone component is required, for instance in order to build a compact model of the component [LKBB12]. In such a case a large number of transformers with different geometries should be fabricated and measured, depending on the parameter range to be covered. Other common application scenarios are the verification of EM simulation tools or the debug of circuits. For this purpose a certain number of test structures are needed: a main structure, in which the four terminals of the transformer are routed to two sets of Ground-Signal-Ground-Signal-Ground (GSGSG) pads, plus a few de-embedding structures, which depend on the utilized de-embedding methodology. In order to achieve the desired accuracy, a suitable choice of de-embedding methdology [Lou16] and a proper design of the corresponding structures are crucial. This tends to become more of an issue as the operating frequency of the circuit grows into the mmW range, since the smaller device size required to obtain a sufficiently large SRF makes it more sensitive to the non-idealities of the de-embedding structures. In the study presented in this section the main and de-embedding structures are EM-simulated to predict the accuracy of the de-embedded results. In this way some layout

guidelines are derived which allow to obtain good accuracy of the extracted parameters in spite of the simple de-embedding method. Several test structures implementing these guidelines are fabricated on a testchip in the 16FF process and measured to demonstrate the concept.

5.5.1 Deembedding Methodology

In this study a very simple two-step open/thru de-embedding method [GTR+07] is employed. Besides the main test structure (Figure 5.13a), this technique requires the open (Figure 5.13b) and the so-called "loop-back" thru (Figure 5.13c). The open is obtained by removing the DUT from the main structure, while the loopback thru is obtained from the open by shorting the end points of each pair of neighboring feedlines. This method belongs to the category of lumped methods discussed in the introduction to Chapter 3 and, as such, suffers from the same limitations.

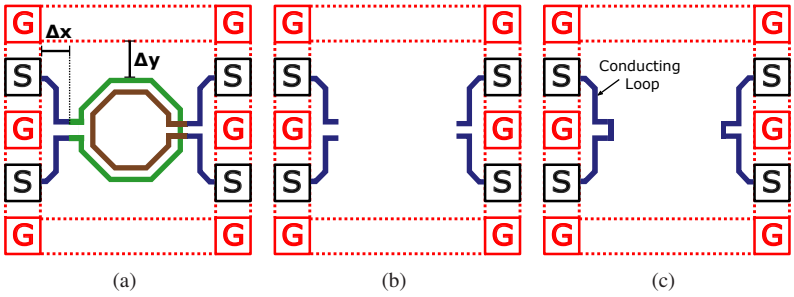


Figure 5.13: Schematic representations of the measurement structures: (a) main, (b) open and (c) loop-back thru (reprinted from [2] ©2022 IEEE).

This work proposes a simulation-based methodology to assess the best way to lay out the test structures prior to hardware fabrication. The goal is to achieve the highest possible accuracy in the extraction of the transformer parameters, which can be computed starting from the simplified model of Figure 5.4b [LKBB09]:

$$L_p = \frac{\mathrm{Im}(Z_{dd,11})}{\omega} \tag{5.16}$$

$$L_s = \frac{\mathrm{Im}(Z_{dd,22})}{\omega} \tag{5.17}$$

$$k_m = \sqrt{\frac{\mathrm{Im}(Z_{dd,21})\,\mathrm{Im}(Z_{dd,12})}{\mathrm{Im}(Z_{dd,11})\,\mathrm{Im}(Z_{dd,22})}} \tag{5.18}$$

$$R_p = \mathrm{Re}(Z_{dd,11}) \tag{5.19}$$

$$R_s = \mathrm{Re}(Z_{dd,22}) \tag{5.20}$$

where $\mathbf{Z_{dd}}$ is the two-port differential-mode impedance matrix of the transformer. The key concept is to simulate the main and de-embedding structures, including the pads, using a commercial full-wave EM solver and compute the de-embedded parameters using the equations below:

$$\mathbf{Y_{to}} = \mathbf{Y_{thru}} - \mathbf{Y_{open}} \tag{5.21a}$$

$$\mathbf{Z_{tr,de}} = (\mathbf{Y_{main}} - \mathbf{Y_{open}})^{-1} \tag{5.21b}$$

$$Z_{tr,de}(i,i) = Z_{tr,de}(i,i) - 0.5 * Y_{to}^{-1}(i,i) \quad (i = 1\ldots4) \tag{5.21c}$$

where $\mathbf{Y_{main}}$, $\mathbf{Y_{open}}$ and $\mathbf{Y_{thru}}$ are the four-port admittance matrices of the main, open and thru structures respectively and $\mathbf{Z_{tr,de}}$ is the four-port impedance matrix of the de-embedded transformer. In this way the measurement procedure is closely reproduced in simulation. These results are then compared to those obtained simulating the transformer standalone ($\mathbf{Z_{tr,sa}}$) to evaluate the de-embedding error. A very useful way to visualize the performance of a certain test structure is to plot the percent de-embedding errors ΔL_p, ΔL_s, ΔR_p, ΔR_s and Δk_m as a function of frequency (see Figure 5.14). For instance to calculate ΔL_p, the de-embedded $L_{p,de}$ and the standalone $L_{p,sa}$ are first computed from

the differential parts of $\mathbf{Z_{tr,de}}$ and $\mathbf{Z_{tr,sa}}$ respectively using (5.16) and finally ΔL_p is computed as:

$$\Delta L_p = \frac{L_{p,de} - L_{p,sa}}{L_{p,sa}} \tag{5.22}$$

The other de-embedding errors are obtained in a similar way.

5.5.2 Layout Considerations for the Test Structures

Applying this EM simulation-based method one can easily verify that the two layout features which mostly affect the de-embedding accuracy are the distance of the DUT from the pad frame and the area of the loop formed by the feedlines. Table 5.3 shows that ΔL_p decreases in absolute value as the distances Δx and Δy (see Figure 5.13a) between the DUT and the pad frame increase, since the unwanted magnetic coupling between the DUT and the ground structure is reduced. The first layout guideline is therefore to place the DUT far enough from the pad structures to make this effect negligible.

Table 5.3: Simulated de-embedding error on inductance, resistance and coupling factor at 10 GHz for different values of Δx and Δy (reprinted from [2] ©2022 IEEE).

Δx (µm)	Δy (µm)	ΔL_p (%)	ΔL_s (%)	Δk_m (%)	ΔR_p (%)	ΔR_s (%)
34.5	17.6	-16.6	-16.4	-3.2	-6.6	-5.4
34.5	41.6	-11.8	-11.8	-1.9	-8.2	-6.7
60.9	41.6	-6.4	-6	-2.3	-3.6	-1.7

The second critical effect is the over de-embedding of the inductance. It occurs because the thru structure forms a conducting loop which is not present in the original structure (see Figure 5.13c), and has therefore slightly higher inductance than the one which should be de-embedded. In order to mitigate this effect the feed lines have to be laid out in such a way to minimize the area of the loop in the thru structure. Table 5.4 shows that using the layout in Figure 5.15b instead of that in Figure 5.15a the inductance and coupling factor de-embedding errors improve from -13.2% to -2.6% and from -10% to -3.4% respectively, but the resistance de-embedding error degrades from -1.3% to -8.3%. The reason

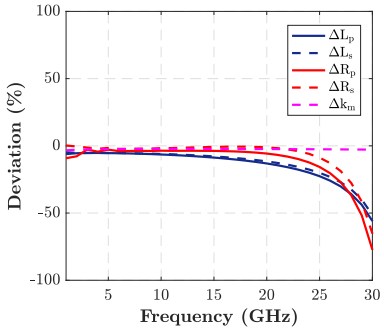


Figure 5.14: Simulated de-embedding errors for Δx = 60.9 µm and Δy = 41.6 µm (reprinted from [2] ©2022 IEEE).

is probably that the feed lines run close to the ground plane and to one another for a significant length, which gives rise to substantial coupling and violates the fundamental assumptions of the de-embedding methodology. A good trade-off between the inductance/coupling factor and resistance de-embedding accuracy can be attained using the feedline layout shown in 5.15c, which is taken as optimal solution.

Table 5.4: De-embedding error on inductance, resistance and coupling factor at 10 GHz for different feed line layouts (reprinted from [2] ©2022 IEEE).

Feed-line	**ΔL_p (%)**	**ΔL_s (%)**	**ΔR_p (%)**	**ΔR_s (%)**	**Δk_m (%)**
(Figure 5.15a)	-13.2	-12.7	-0.8	-1.3	-10
(Figure 5.15b)	-2.6	-2.8	-8.9	-8.3	-3.4
(Figure 5.15c)	-7.1	-6.2	-5.1	-4.2	-1.4

5.5.3 Measurement Results

A testchip with six different transformer layouts employing the optimized de-embedding structures described in section 5.5.2 has been designed and fabricated in the 16FF process, as shown by the die micrograph in Figure 5.16. These devices utilize the two top copper layers of the stack, M8 and M9 and the aluminum layer AL (see Figure 2.12a). The geometrical features of the various

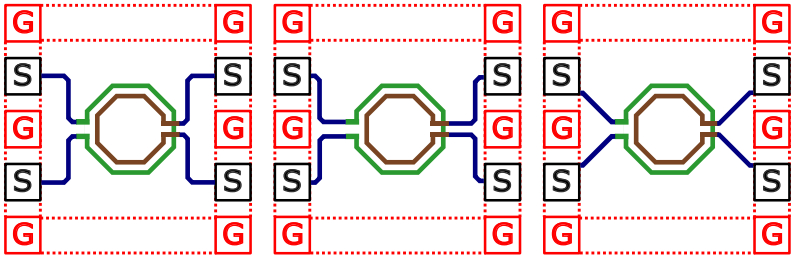


Figure 5.15: Test structures with different feed line layouts: (a) large-loop feed lines, (b) small-loop feed lines and (c) 45° feed lines (reprinted from [2] ©2022 IEEE).

DUTs are reported in Table 5.5, where the transformer parameters are defined as in section 5.4. Thanks to their relatively small size all the transformers show SRF above 80 GHz.

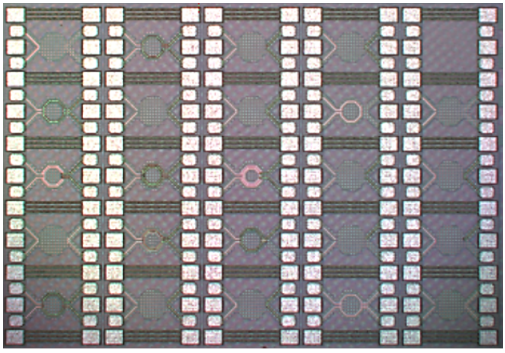


Figure 5.16: Die micrograph (reprinted from [2] ©2022 IEEE).

The 4-port measurements have been performed through on-wafer probing using a VNA calibrated up to the probe tips with frequency range from DC to 67 GHz. Figure 5.17 shows the simulated and measured de-embedding error on inductance (Figure 5.17a), resistance (Figure 5.17b) and coupling factor (Figure 5.17c) for DUT1. The measured de-embedding error is computed in the exact same way as shown in section 5.5.1, but using measured instead of simulated

Table 5.5: Geometrical features of the transformers in the testchip (reprinted from [2] ©2022 IEEE).

DUT	Type	M_p	M_s	$N_{t,p}$	$N_{t,s}$	r_p (µm)	r_s (µm)	w_p (µm)	w_s (µm)
1	Stacked	M9	AL	1	1	35	35	4	4
2	Stacked	M9	M8	1	1	35	35	4	4
3	Stacked	M9	AL	2	1	35	31	4	4
4	Interleaved	M9	M8	1	1	35	30	4	4
5	Interleaved	AL	AL	1	1	29	35	4	4
6	Interleaved	M9	M8	2	1	35	30	4	4

data. Since it is inherently not possible to measure the standalone transformer, the error is computed with respect to the simulated data:

$$\Delta L_{p,meas} = \frac{L_{p,de,meas} - L_{p,sa,sim}}{L_{p,sa,sim}} \tag{5.23}$$

The measured de-embedding errors of inductance and coupling factor show very good agreement with the simulations. The resistance instead shows quite substantial deviation due to the low measured values (~1 Ω), resulting in large relative error.
Table 5.6 reports the measured (subscript "m") and simulated (subscript "s") values of the de-embedding errors at 40 GHz for the remaining devices. The data show that the measured ΔL_p and Δk_m are generally in good agreement with the simulated values. The residual disagreement stems from the fact that the EM simulation of the pad structure requires significant layout simplification in order to complete in a reasonable time, which comes invariably at the cost of lower accuracy. Also in this case significant ΔR_p and ΔR_s are observed. The overall satisfactory agreement between measurements and simulations validates the discussed layout guidelines for the de-embedding structures.

5.6 Summary

This Chapter has analyzed different techniques to determine the insertion loss caused by the impedance MNWs within a PA without requiring the simulation

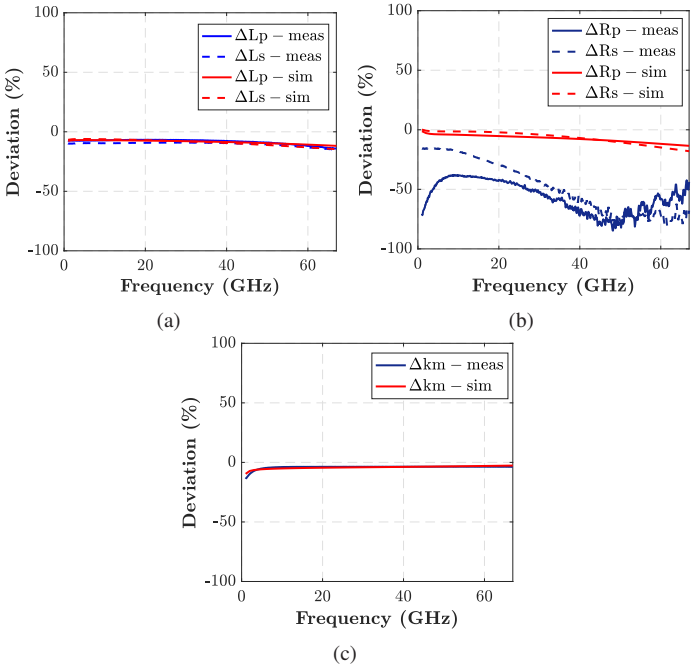


Figure 5.17: Measured and simulated de-embedding error on (a) inductance, (b) resistance and (c) coupling factor for DUT1 (reprinted from [2] ©2022 IEEE).

of the complete circuit. Building on this knowledge, a large-signal generalization of the transducer gain has been identified as the most appropriate figure of merit for the design of the matching networks. A layout optimization technique based on this figure of merit was presented and applied to the synthesis of transformer-based MNWs with minimum loss for the 16FF and 22SOI output stages designed in Chapter 4. This analysis has shown that transformers operating above the SRF and utilizing the 1:2 turn ratio can effectively reduce the insertion loss in scenarios with large impedance transformation ratios. Furthermore, the methodology allowed for a systematic comparison between the two RF metal stack profiles available in 22SOI. It was demonstrated that the benefit of an additional ultra-thick metal layer lies rather in the reduced interconnect parasitics within the PA core than in a better insertion loss of the transformer. This benefit was shown to be fairly limited for the analyzed PA architecture,

Table 5.6: Measured and simulated de-embedding errors for all DUTs at 40 GHz (reprinted from [2] ©2022 IEEE).

DUT	$\Delta L_{p,m}$ (%)	$\Delta L_{p,s}$ (%)	$\Delta L_{s,m}$ (%)	$\Delta L_{s,s}$ (%)	$\Delta k_{m,m}$ (%)	$\Delta k_{m,s}$ (%)	$\Delta R_{p,m}$ (%)	$\Delta R_{p,s}$ (%)	$\Delta R_{s,m}$ (%)	$\Delta R_{s,s}$ (%)
1	-7.7	-8.6	-9.6	-9.4	-3.6	-3.8	-72.2	-7.7	-58.1	-6.9
2	-5.2	-11.5	-5.8	-10.9	-3	-2.6	-63.4	-18.4	-32.9	-13.4
3	-2.4	-8.3	-6.1	-8.6	-1.6	-2.9	-25.5	-14.6	-37.2	-6.1
4	-4.2	-8.6	-2.3	-7.6	-11.3	-4.4	-60.1	-13.7	-38.9	-7.6
5	-9.1	-7.7	-10.5	-8	-7.8	-6.4	-45.3	-10.5	-34.4	-10.6
6	-4.9	-9.5	-0.5	-9.9	-6.7	-1.6	-59	-23	-36.9	-13.9

which should encourage circuit designers to use the less expensive metal stack profile with only one ultra-thick copper layer. A comparison between the 16FF and 22SOI processes revealed that, contrary to the common understanding, the two technologies show very similar performance. This can be attributed to a large extent to a more carefully optimized RF transistor layout in 16FF. Finally a test structure concept for the characterization of standalone transformers was presented, which can be profitably used for modeling or circuit debug purposes.

6 PA Design and Characterization

The design methodologies presented in Chapters 4 and 5 enabled an unbiased benchmarking and comparison of the 16FF and 22SOI processes. This was done for the design of the output section of a PA at 80 GHz with optimal power-gain-efficiency trade-off. Since the presented conclusions were drawn on the basis of simulations only, they require validation by hardware results. In order to show that the described methodology leads to a real and functional design, a multi-stage PA should be used as circuit demonstrator. This is indeed the typical architecture utilized in mmW transceiver due to the limited gain of the active devices at such high frequencies. Besides reflecting the simulated behavior, the design should also be free of unwanted oscillations, which are a common plague in analog circuits with high gain.
Serving this validation purpose, this Chapter deals with the design and characterization of full PA prototypes in the 16FF and 22SOI processes. The utilized ouput stages and output MNWs are those designed in the previous two Chapters, with only some minor differences dictated by external constraints. Besides the functional circuit design, lot of attention is devoted to the modeling techniques for best model-hardware correlation and the necessary checks to detect undesired oscillations prior to IC fabrication. On the measurement methodology front, a suitable technique is described to deal with the challenge of on-wafer fully differential E-band measurements. In the last part of the Chapter, the measured performances of the 16FF and 22SOI prototypes are analyzed and compared to previous art. The main results related to the 16FF PA prototype presented in this Chapter are extracted from [3].

6.1 PA Prototype in 16FF

6.1.1 Design Considerations

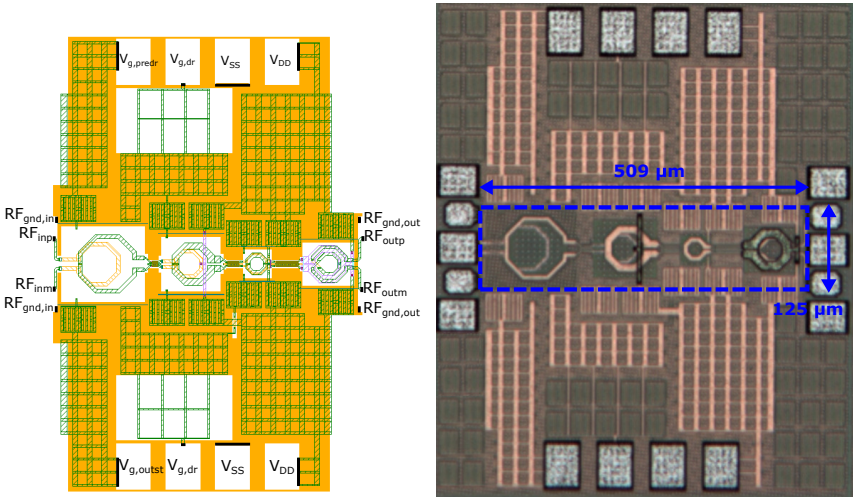
In mmW circuit design it is common practice to exploit the input and output transformers to convert the differential signals into single-ended. While this is sometimes justified by the application, for instance if the output of the PA drives directly the antenna of the TX, in many cases the main motivation is to provide a more measurement-friendly interface. As discussed in section 5.4, the design of a balun comes with imbalance requirements which typically restrict the design space to transformers which operate below the SRF, leading to performance degradation. For these reasons in this work it is decided to use a fully differential configuration for the PA prototypes, with the input and output routed to GSGSG pads. This simplifies the design of the input and output transformers but introduces new challenges in the measurement, which has to be performed with a 4-port VNA in true differential mode. In order to measure at frequencies up to 110 GHz, the utilized VNA model requires a frequency extender at each port. Due to the high loss of the cables and of the extenders at such high frequencies, the VNA can provide at most −3 dBm differential power at the input of the PA. With this set of boundary conditions it is found that in 16FF two driver stages are required to make sure that the output stage attains saturation at the design frequency of 80 GHz.
As mentioned in section 2.1.4, the design starts from the output stage and proceeds backwards. The output stage is the one determined in section 4.3.1, with the only difference that the input signal lines run on M8 instead of M9. This reflects a previous version of the core design with reduced C_{gd} which was in use at the time of the tape-out, as briefly mentioned in section 4.1. The output MNW is obtained starting from the one with M = 8 in section 5.4.1, which is re-tuned replacing $Z_L = 100\,\Omega$ with a value which includes the contributions of the output RF pads and the attached feedlines. The driver stages utilize the same unit NDP cell as the core. Since the combined gain of the core and the output MNW is less than 10 dB, PAE_{PA} is significantly affected by PAE_{dr} (2.6), where the subscript "dr" stands for "driver". The multiplicity $M_{dr} = 4$ is chosen in such a way that the driver is in 0.5-dB compression at the output power level which forces the output stage in 3-dB compression. This ensures that the driver does not compress before the output stage [YMYZ15], while simultaneously allowing to maintain decent PAE_{dr}. For the same reason, the gate bias $V_{g,dr}$

is set slightly lower than in the output stage, namely $V_{g,dr} = 0.45\,V$. The pre-driver, abbreviated "predr" in the formulas, is designed to be approximately in 0.2-dB compression at the output power level which forces the driver in 0.5-dB compression. This is achieved with $M_{predr} = 2$ and $V_{g,dr} = 0.5\,V$. Since the driver and output stage have a combined gain of more than 18 dB, PAE_{PA} is almost insensitive on PAE_{predr}. The inter-stage and input MNWs are also designed using the layout optimization methodology of section 5.4. Table 6.1 shows a list of all the transformer-based MNWs in the PA with the related geometrical features, as defined in section 5.4.

Table 6.1: List of transformer-based MNWs used in the 16FF PA prototype.

Type	M_p / M_s	n_p:n_s	r_p / r_s (µm)	w_p / w_s (µm)	s_p / s_s (µm)	x_{off} (µm)
Input MNW	M8 / M9	1:1	36 / 48	6 / 6	- / -	15
Inter-stage MNW 1	M8 / M9	1:1	23 / 35	5 / 10.8	- / -	8
Inter-stage MNW 2	M8 / M9	1:1	17 / 18	7 / 2	- / -	0
Output MNW	AL / M9	1:2	31 / 30	12 / 5	- / 6	4

The complete PA includes two sets of four horizontally-arranged DC pads, located respectively at the top and bottom side of the die, as shown in Figure 6.1b. They are used to set the DC ground V_{SS} and to provide the active devices with the gate bias voltages $V_{g,predr}$, $V_{g,dr}$ and $V_{g,outst}$ and the supply voltage V_{DD} through the center taps of the transformers. Since a minimum vertical separation of 150 µm is required between the DC and RF pads to be able to connect the DC and RF probes at the same time, a large chip area remains potentially empty. This is filled with large decoupling capacitors, which stabilize the supply voltage and improve the Common Mode (CM) stability, as discussed at length in section 6.1.2. The PA includes also a low-resistivity ground mesh on metal layers M1-M8 to minimize the IR drop. This is only critical for the DC currents, since in a differential circuit negligible signal current flows through the ground. The ground mesh serves also the purpose of connecting the grounds of the supply and of the RF probes, bringing them to the same potentials. The die micrograph of Figure 6.1b shows that the prototype occupies an area of $0.064\,mm^2$, excluding the pads and the decoupling network.
For the simulation of the complete circuit, a comprehensive representation is built using the PDK models for the transistors and the capacitors, along with an S-parameters model for the passives and the interconnects. This is accomplished using an EM simulator which allows to exclude the PDK components from the



(a) (b)

Figure 6.1: (a) Simplified layout for EM simulation and (b) die micrograph of the three-stage 16FF PA prototype.

simulation, generate the S-parameter model of the interconnects and finally assemble the circuit elements together. In order to avoid unacceptably long simulation times, the RF and DC pad structures and the decoupling capacitors are excluded from the EM simulation. This is shown in the simplified layout of Figure 6.1a, in which the interface ports of the model are highlighted in black. The excluded elements are taken into account separately in the top-level simulation model of Figure 6.2, where the block "PA" is the comprehensive representation of the PA discussed above. The RF pad structure is modeled by two ideal lumped capacitors of capacitance C_{pad} connected between each RF input/output and the corresponding RF ground. The value $C_{pad} = 25\,fF$ is determined from an EM simulation of the pad structure standalone. The decoupling capacitors C_{dec} are also placed externally as lumped components, whereas the DC pads are not modeled at all due to their negligible impact on the performance.

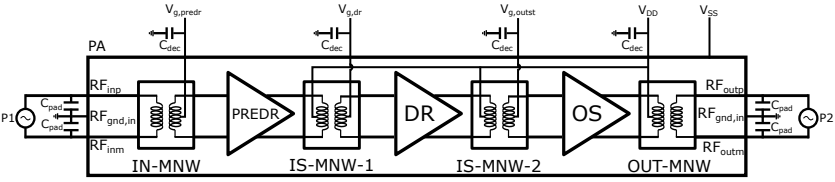


Figure 6.2: Top-level simulation model of the three-stage PA prototype.

6.1.2 Stability Analysis

An unstable behavior of a circuit is defined as the unwanted build-up of oscillations caused by the presence of a loop with positive feedback across a block with large gain. In RF and mmW amplifiers utilizing the CS configuration this unwanted path typically goes through the C_{gd} of the active device. Stability is particularly critical in differential PAs because oscillations can potentially build up in the Differential Mode (DM) as well as in the CM, especially at low frequencies, at which the gain of the active devices is large. As a consequence, stability has to be carefully checked for both modes from DC up to the operating frequency. One advantage of differential PAs based on the NDP is that in-band DM stability is obtained by design, but unfortunately neither out-of-band DM stability nor CM stability are guaranteed [DR14]. Over the years several methods have been proposed to determine potential oscillations prior to the fabrication of the IC [Sua15]. Unfortunately none of these methods is "universal" and in fact each of them proves more or less effective depending on the type of instability, which is typically not known a priori. One popular method consists in ckecking that the stability factor μ (6.1) for the desired mode, that is DM or CM, be greater than 1 across the entire frequency range.

$$\mu = \frac{1 - |S_{11}|^2}{\left|S_{22} - S_{11}^{*}\Delta S\right| + |S_{12}S_{21}|} > 1 \tag{6.1}$$

In order to emulate the variation of the bias point over time in Large Signal (LS) operation, the gate and drain bias voltages of all stages are swept in the ranges $[0,2V_{gs}]$ and $[0,2V_{DD}]$ respectively to cover all the theoretically possible combinations. The CM stability can be assessed using the simulation setup in Figure 6.3 to derive the stability factors $\mu_{cc,i}$, where the subscript i refers to

the stage under analysis (i = 1, 2, 3). For each stage the two single-ended ports for the S-parameters (SP) analysis are connected to the $V_{g,i}$ and V_{DD} lines respectively. The input and output of the PA are terminated on 100 Ω differential to reflect the actual operation of the circuit. The DM stability can be assessed from μ_{dd}, which is obtained from the test setup of Figure 6.2, with two differential ports placed at the main input and output of the circuit. To be completely rigorous, this analysis should be also performed for each stage individually, but is unfortunately not possible due to unavailability of the internal signal nodes in the PA model.

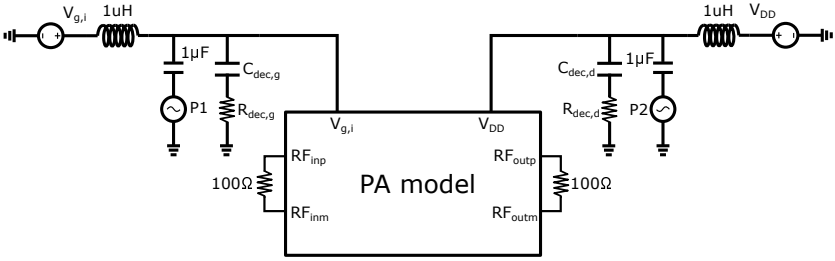


Figure 6.3: Test setup for the CM stability analysis.

For the CM analysis it is important to model the supply network as accurately as possible. The parasitic inductance of the DC cables used for the measurement is particularly critical as it can favor the build-up of CM oscillations. In the model of Figure 6.3 it is set to a conservative estimate of 1 µH. The decoupling capacitors $C_{dec,g}$ and $C_{dec,d}$ are used to damp any potential CM oscillation and for this reason they come with the additional series resistances $R_{dec,g}$ and $R_{dec,d}$ to lower the Q-factor. Even though they are located on-chip, in this testbench they are added as external components for the reasons explained in section 6.1.1. The effect of C_{dec} on the CM stability of the output stage is demonstrated in Figures 6.4a and 6.4b: without any decoupling capacitors there is a potential CM instability ($\mu_{cc} < 1$) around 1 GHz, whereas with decoupling capacitors unconditional stability ($\mu_{cc} > 1$) is attained over the entire frequency range. The utilized capacitors are a combination of MOM and Metal-Insulator-Metal (MIM) devices with total capacitances of $C_{dec,g} = 104.7$ pF and $C_{dec,d} = 348.4$ pF. The associated series resistances are $R_{dec,g} = 0.36\,\Omega$ and $R_{dec,d} = 0.14\,\Omega$. Figures 6.5a and 6.5b show that the pre-driver and driver stages are also unconditionally stable for the CM with the same amount of decoupling capacitance as in the

output stage. Finally, Figure 6.6 shows that the DM is stable over the whole frequency range.

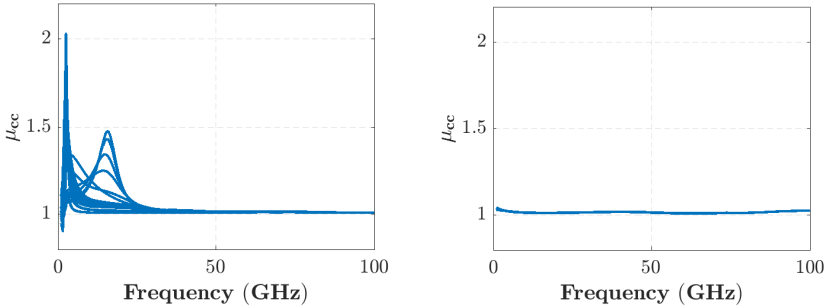


Figure 6.4: Common-mode stability factor μ_{cc} of the 16FF PA prototype as a function of frequency for different bias conditions, with excitation applied to the output stage, (a) without and (b) with decoupling capacitors.

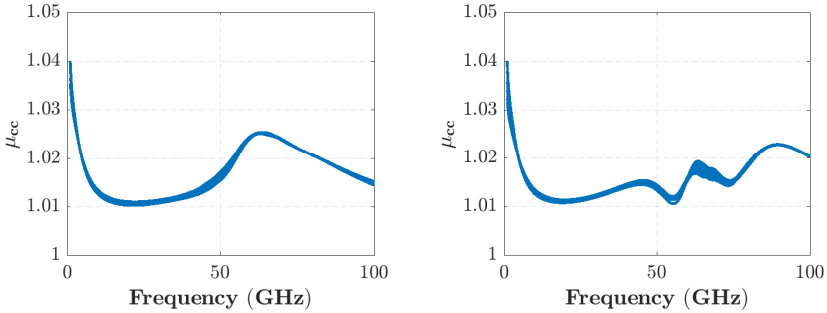


Figure 6.5: Common-mode stability factor μ_{cc} of the 16FF PA prototype as a function of frequency for different bias conditions, with excitation applied to (a) the pre-driver and (b) the driver.

The µ-factor analysis is based on the S-parameters and as such is only suitable to detect linear instabilities [MNQ+99]. There exists another class of instabilities called parametric oscillations, which can arise in the large-signal regime due

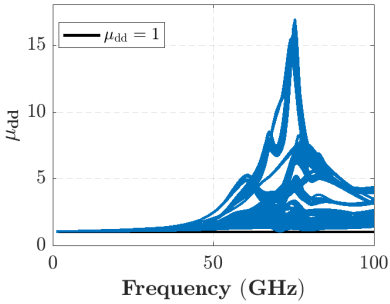


Figure 6.6: Differential-mode stability factor μ_{dd} of the 16FF PA prototype as a function of frequency for different bias conditions.

to the non-linearity of the parasitic capacitances of the active devices. While a rigorous analysis of these phenomena [SJR06] is beyond the scope of this work, one simple way to detect them is to run a transient analysis injecting a short pulse with large amplitude and broad frequency content into the PA and checking for oscillations at the output. The test is run for the same scenarios considered in the μ-factor analysis, but in this case the voltage waveforms $v_{out,p}$ and $v_{out,m}$ at the two differential output terminals are sampled. The injected signal is a 1.6 V triangular pulse with rise and fall times of 1 ps each. As shown in Figure 6.7, the output response dies out in about 0.2 ns in all the analyzed scenarios, which indicates a stable behavior.

6.1.3 Measurement Results

The three-stage 16FF prototype was fabricated and characterized using a 4-port VNA in true differential mode calibrated up to the probe tips. In the reported sample the V_{gs} bias voltages of the three stages were increased by 3.5 mV to obtain the same quiescent drain current as in the simulation, $I_d = 134$ mA. This step is necessary to compensate for small fluctuations of the transistor V_t caused by process variations. The resistive losses on the supply line caused by the cables and the DC probes were compensated by a suitable increase of V_{DD} and reduction of V_{SS} to ensure a potential difference of exactly 0.8 V between the V_{DD} and V_{SS} pads. The measured and simulated differential-mode

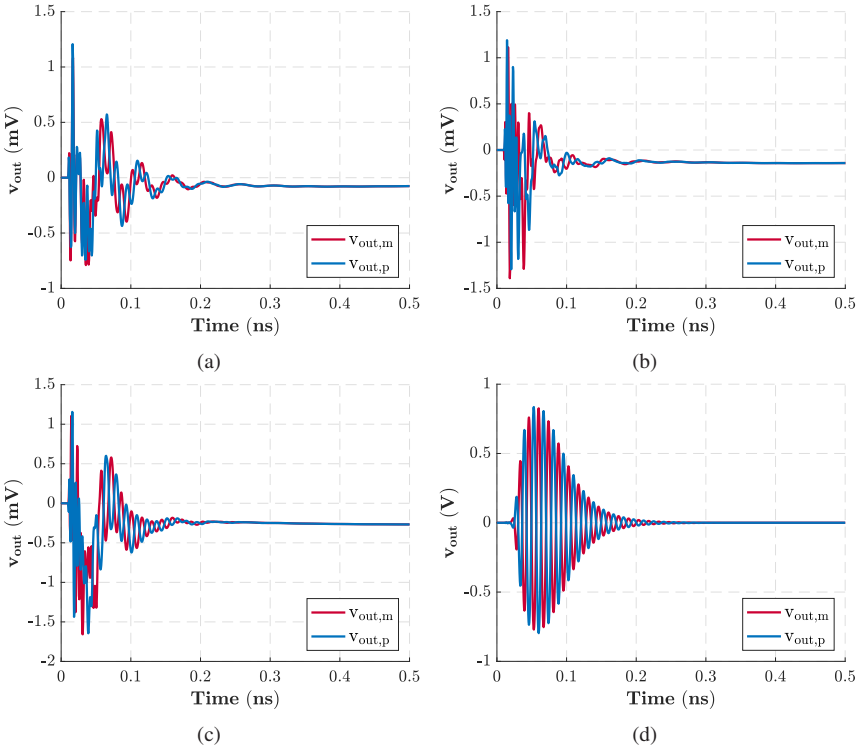


Figure 6.7: Transient stability analysis of the 16FF PA prototype with CM step signal applied on (a) pre-driver, (b) driver and (c) output stage and (d) with DM step signal applied to the input.

small-signal S-parameters $|S_{dd,ij}|$ are plotted in Figure 6.8, showing excellent correlation for $|S_{dd,21}|$ and good correlation for $|S_{dd,12}|$. As for $|S_{dd,11}|$ and $|S_{dd,12}|$, some deviation is observed due to the approximated model used for the RF pads. In these simulations the layout parasitics of the amplifying stages were modeled using RC extraction up to M3 and EM simulations from M3 to M9, which resulted in better model-hardware correlation compared to the full EM simulation methodology discussed in section 4.1. Since the substrate contact of the NDP was drawn manually, as explained in section 4.3.1, RC extraction is required to generate a model of the substrate network. The peak of $|S_{dd,21}|$

is located at 70 GHz, about 10 GHz below the design frequency. This effect is a consequence of the utilized design methodology and becomes apparent upon incorporation of the driver stages.

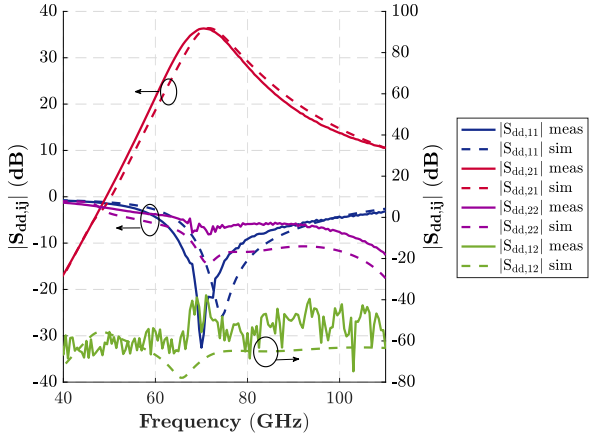


Figure 6.8: Measured and simulated differential-mode S-parameters of the three-stage 16FF PA prototype.

The large-signal measurements have been performed using the same setup, with the four VNA sources and receivers leveled from −43 dBm to −6 dBm, so that a true differential signal with power P_{in} ranging from −40 dBm to −3 dBm could be presented at the input of the PA. The VNA measures the differential G of the PA, which is equal to the large-signal $|S_{dd,21}|$, and uses it to derive P_{out}, taking into account the insertion loss of the input RF probe. Measuring the current consumption I_d with the power supply, the PAE can be easily determined. Figure 6.9 shows that the measured P_{sat} and PAE_{peak} as a function of frequency match very well to the simulations. The slight deviation in the low-frequency range can be explained by the mismatch between the measured and simulated $|S_{dd,11}|$ and $|S_{dd,22}|$. The largest FoM_{3dB} is achieved at 70 GHz, which coincides with the peak of $|S_{dd,21}|$ observed in Figure 6.8. At this frequency the PA has G_{ss} = 34.9 dB, P_{sat} = 15.2 dBm and PAE_{peak} = 30.3%, as shown in Figure 6.10a. The measured I_d in Figure 6.10b shows very good agreement with the simulations and a typical class-A behavior, with the active stages fully turned on also at very low P_{in} levels.

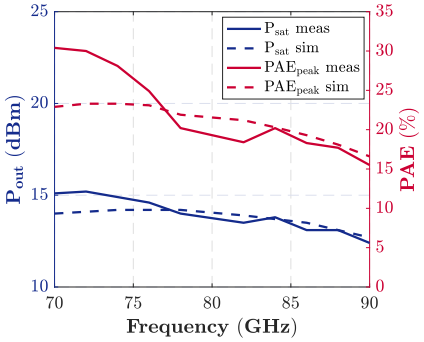


Figure 6.9: Measured and simulated P_{sat} and PAE_{peak} as a function of frequency of the 16FF PA prototype.

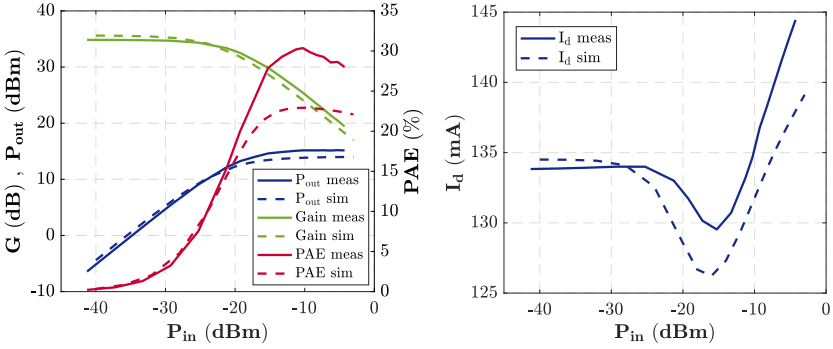


Figure 6.10: Measured and simulated (a) P_{out}, G, PAE and (b) DC current consumption of the 16FF PA prototype as a function of P_{in} at 70 GHz.

6.2 PA Prototype in 22SOI

Most of the design and the measurement considerations presented in section 6.1 are applicable also to the PA prototype in 22SOI. However, due to time constraints and export control issues, the circuit was designed for a different frequency and output power level. Specifically, the center frequency was chosen to be 70 GHz instead of 80 GHz to comply with the US export control rules.

Moreover, the multiplicity of the output stage was set to the conservative value $M_{outst} = 4$, as the results presented in section 4.3.2 were not available at the time of the tape-out. Finally, similarly to 16FF, the input lines of the core were drawn on M9 instead of M10. Consistently with the conclusions of section 4.3.2, the gate bias voltage of the output stage was set to $V_{g,outst} = 0.5\,V$. Due to the lower W_{tot} and operating frequency compared to 16FF, a single driver stage was found to be sufficient to saturate the circuit at the center frequency. The driver multiplicity and bias voltages were chosen to be $M_{dr} = 1$ and $V_{g,dr} = 0.5\,V$ to ensure 0.5-dB compression at the output power level which drives the core stage in 3-dB compression. The features of the input, output and inter-stage transformer-based MNWs obtained from the layout optimization methodology of section 5.4 are reported in Table 6.2. The prototype occupies an area of $0.083\,mm^2$ excluding the measurement pads and the decoupling network, as shown in the die micrograph in Figure 6.11.

Table 6.2: List of transformer-based MNWs used in the 22SOI PA prototype.

Type	M_p / M_s	n_p:n_s	r_p / r_s (µm)	w_p / w_s (µm)	s_p / s_s (µm)	x_{off} (µm)
Input MNW	M10 / M9	1:1	62 / 58	8 / 4	- / -	20
Inter-stage MNW	M10 / M9	1:1	28 / 32	6 / 12	- / -	4
Output MNW	M10 / M9	1:2	42 / 44	16 / 6	- / 7	4

The measured small-signal S-parameters in Figure 6.12a show that $|S_{dd,21}|$ peaks at about 64 GHz. This is significantly below the design frequency, but the shift of 6 GHz is lower than the one observed in 16FF, which seems to be a consequence of the lower number of stages. A sharp drop of $|S_{dd,21}|$ around 35 GHz is also observed, which results in a discrepancy of approximately 5 dB with the simulations at frequencies above this value. The reason turns out to be that the circuit has an undesired oscillation at about 17 GHz, even without any RF excitation applied. This is clearly visible from the output spectrum in Figure 6.12b, where OUT+ and OUT- represent the two output terminals of the PA. This effect was not predicted by the stability analysis performed before tape-out due to some inaccurate assumptions in the testbench. A µ-factor analysis conducted a posteriori with a more accurate testbench shows indeed a potential CM instability around 20 GHz (see Figures 6.13a and 6.13b), which is very close to the measured oscillation. The simulation also reveals a potential CM as well as DM oscillation around 40 GHz, as shown in Figures 6.13b and 6.13c,

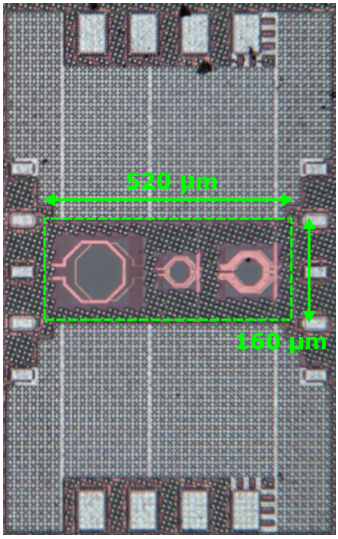


Figure 6.11: Die micrograph of the two-stage 22SOI PA prototype.

which is however not observed in the measurement. Applying an in-band CW excitation at the input, the oscillation decreases in amplitude as the power of the input signal increases and eventually it disappears completely, as shown in Figures 6.14a and 6.14b. The reason is that at large signal levels the gain of the PA gets compressed and eventually the loop gain of the system is no longer sufficient for the onset of an oscillation. In the CW measurements as a function of P_{in} this effect translates into a progressive reduction of the gap between the measured and simulated gain with increasing P_{in}, as shown in Figure 6.15a. The measurement was taken at $f_0 = 66$ GHz, at which the PA shows the largest FoM. The large-signal performance as a function of frequency (Figure 6.15b) shows also very good correlation between model and hardware. The simulations were performed using a model of the PA with the layout parasitics extracted entirely by means of EM simulations, without the need of RC extraction below M3. The reason is that, differently from 16FF, the substrate contact from the standard transistor PCell was used, so that the effect of the substrate network is correctly captured within the transistor model. Based on the collected data, it is expected that the excellent correlation between simulations and measurements observed in large-signal conditions will be extended to the small-signal regime after

applying suitable stabilization measures. For this reason the demonstrated PA prototype validates the conclusions about 22SOI presented in Chapters 4 and 5.

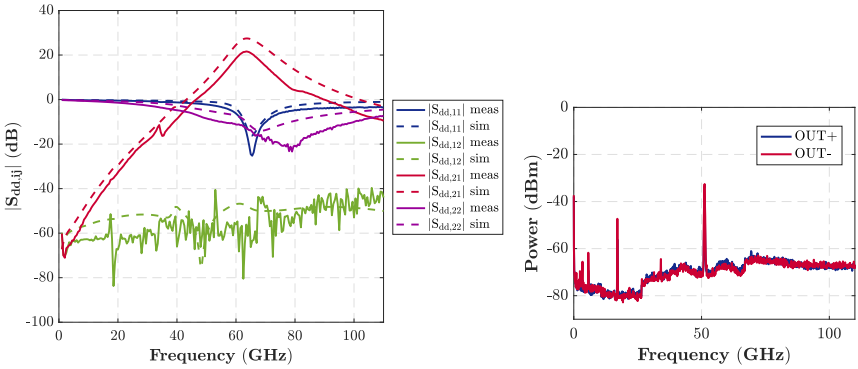


Figure 6.12: (a) Measured and simulated differential-mode S-parameters and (b) output spectrum with DC bias an no RF excitation of the two-stage PA prototype in 22SOI.

6.3 Comparison to Previous Art

Table 6.3 shows a comparison with previous art in the E-band in various technologies. For this comparison a modified version FoM_{mod} of the ITRS figure of merit is used, with G_{ss} divided by N_{stages} to limit the impact of the gain, as shown in the footnote of Table 6.3. The displayed data confirms the well-known fact that FinFET technologies [CPH19, CCW+21] show worse performance compared to SOI [CCE20], bulk CMOS [ZR15], SiGe [WR19] and III-V semiconductor technologies [GURP15] for the implementation of mmW PAs.

The 16FF prototype achieves competitive FoM_{mod} with previous art in FinFET technologies [CCW+21] in spite of the simpler architecture and the lower V_{DD}. The 3dB-bandwidth BW is only 9 GHz because the impedance matching has been performed only at the center frequency. The core area of 0.064 mm^2 is in line with those reported by other works in deeply scaled CMOS technologies.

As far as the 22SOI prototype is concerned, the measured FoM_{mod} is signif-

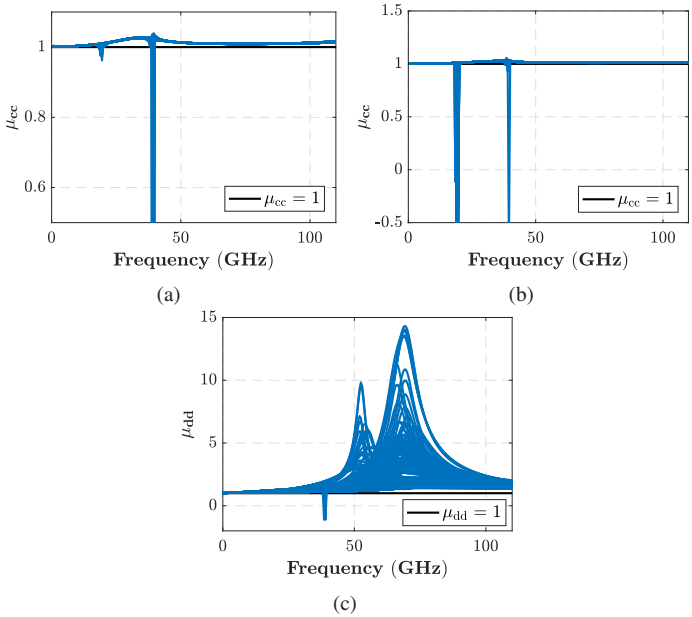


Figure 6.13: µ-factor stability analysis for the 22SOI PA prototype for different bias conditions (a) with CM excitation on driver, (b) with CM excitation on output stage, (c) with DM excitation.

icantly lower than the one reported in [CCE20]. This primarily stems from not utilizing the stacked-FET architecture, thereby sacrificing a key benefit of SOI technologies (refer to section 2.3). Coupled with the sub-optimal design choices discussed in section 6.2, the resulting P_{out} falls short of the potentially achievable value. The core area of 0.083 mm^2 is larger than that of the 16FF prototype in spite of the smaller number of stages, mostly due to the large size of the input MNW.

6.4 Summary

This Chapter discussed the design of the complete PA chain starting from the optimized active stages and matching networks determined in Chapters 4

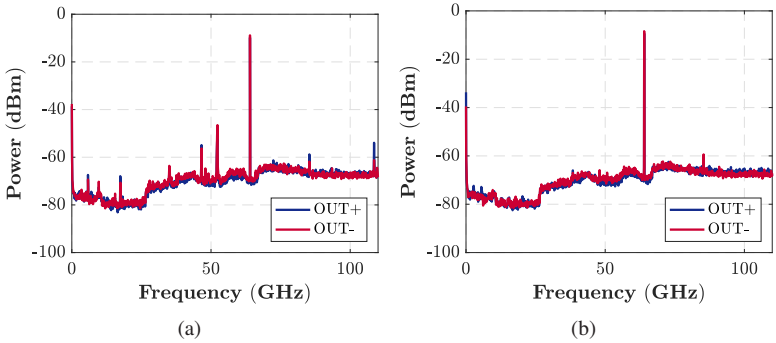


Figure 6.14: Output spetcrum of the 22SOI PA with CW input differential signal at 64 GHz with (a) $P_{in} = -16$ dBm and (b) $P_{in} = -14$ dBm.

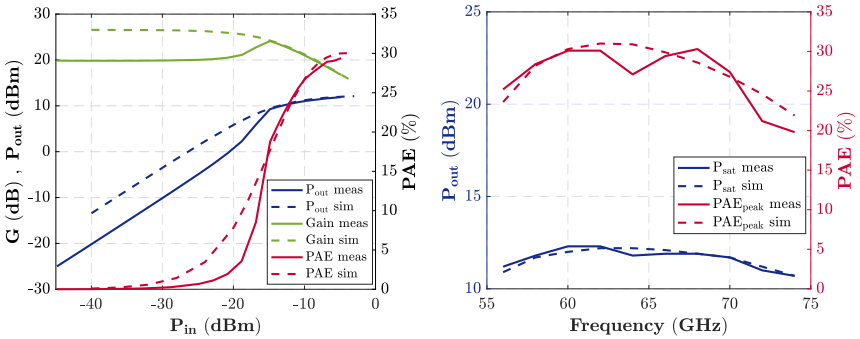


Figure 6.15: Measured and simulated large-signal performance of the 22SOI PA prototype: (a) P_{out}, G and PAE vs P_{in} at 66 GHz and (b) P_{sat} and PAE_{peak} vs frequency.

and 5 respectively. Techniques for the modeling of the PA as well as for the assessment of the stability were also discussed. Following these guidelines, one PA prototype with three stages and center frequency of 80 GHz and one with two stages and center frequency of 70 GHz were fabricated in the 16FF and 22SOI processes respectively. The 16FF prototype demonstrated state-of-the-art performance among the E-band PAs in FinFET technologies, thus validating the adopted design methodology. Conversely the performance of the 22SOI prototype remained significantly below the state of the art due to some

Table 6.3: Performance comparison with state-of-the-art E-band PAs

	This work	**This work**	**[CPH19]**	**[CCW+21]**	**[CCE20]**	**[ZR15]**	**[WR19]**	**[GURP15]**
Technology	16 nm FinFET	22 nm FD-SOI	22 nm FinFET	16 nm FinFET	22 nm FD-SOI	40 nm bulk CMOS	0.12 µm SiGe	150 nm InP HBT
Frequency (GHz)	70	66	75	65	64	73	70	74
V_{supply} (V)	0.8	0.8	1	0.95	1.6	1.8	2	1.8
V_{supply} per transistor (V)	0.8	0.8	1	0.95	0.8	0.9	2	1.8
N_{stages} (#)	3	2	2	2	2	2	1	1
Gain (dB)	34.9	26.6 (est.)	16.6	21.4	31	25.3	22	10
P_{sat} (dBm)	15.2	11.9	12.8	17.9	21	22.6	24	26.3
PAE_{peak} (%)	30.3	29.4	26.3	26.5	28.7	19.3	12	27.6
Core Area (mm^2)	0.064	0.083	0.054	0.107	0.0335	0.25	3.34	1.72
BW (GHz)	9	7	24	13	NA	NA	15	33
FoM_{mod} (dB)	78.5	76.3	72.8	79.1	87.2	85.4	93.7	88.1

$$*\mathrm{FoM}_{\mathrm{mod}} = \mathrm{P}_{\mathrm{sat}}(\mathrm{dBm}) + \frac{\mathrm{G}_{\mathrm{ss}}}{\mathrm{N}_{\mathrm{stages}}}(\mathrm{dB}) + 10\log_{10}(\mathrm{PAE}_{\mathrm{peak}}(\%)) + 20\log_{10}(\mathrm{f}_0(\mathrm{GHz}))$$

sub-optimal design choices. Additionally, an unexpected oscillatory behavior was observed in the low frequency range, which disrupted the performance in the small-signal regime. The very good match between simulations and measurements observed for both processes provides a sound validation for the technology considerations made in Chapters 4 and 5, which are also the key conclusions of the entire work.

7 Conclusions

In this research work the suitability of deeply scaled CMOS processes for millimeter-wave PAs was thoroughly analyzed. Key challenges related to the active and passive device modeling and characterization as well as to the circuit design were addressed. The relevance of the gate resistance in FinFET technologies and its impact on the performance of the PA were discussed in detail and a novel measurement technique with only one de-embedding structure was proposed and verified with hardware results in the 16 nm FinFET process. A simple measurement technique for integrated transformers with two de-embedding steps was also described and validated with test structures in 16 nm FinFET.

From the circuit design standpoint one of the main achievements was the development of a technology-independent methodology for the design of E-band PAs with optimal performance. This methodology utilized the ITRS figure of merit and a large-signal generalization of the transducer power gain to optimize the amplifying stages and the transformer-based matching networks respectively. At such high frequencies the fundamental challenge was identified in the design of the output stage, which should be electrically large to obtain the largest possible output power but physically compact to minimize the performance degradation caused by the layout parasitics.

The application of said methodology to the design of PA output sections in the 16 nm FinFET and in a 22 nm FD-SOI processes, coupled with a detailed analysis of the electromigration effects, has lead to numerous interesting conclusions at the technology level. The initial expectation was that 22 nm FD-SOI would show better performance than 16 nm FinFET due to the lower transistor parasitic capacitances, more easily predictable gate resistance, better substrate isolation, lower sensitivity to electromigration and availability of a metal stack option with two ultra-thick metal layers. The two processes have shown in fact very similar performance. This is partly due to the fact that stacked-FET architectures, which are particularly suitable to SOI thanks to the complete substrate isolation, were excluded from this investigation. This was however

done with the precise purpose of focusing on more fundamental features of the two processes. At the RF transistor layout level, the gate pitch was found to play a significant role: it should be large enough to reduce the parasitic capacitance and improve the resilience to electromigration but small enough to prevent an overhead of gate resistance and routing parasitics. In this repect the 16 nm FinFET PDK offers a better trade-off compared to 22 nm FD-SOI, but this could be in principle improved with a custom device layout and model.
As far as the metal stack is concerned, it has been proved that the availability of additional ultra-thick layers brings more benefits in the performance of the amplifying stages than in the insertion loss of the matching networks. In the analyzed frequency range this advantage is relatively modest, so that the metal stack option with only one ultra-thick metal layer allows for a significant cost reduction without major performance penalties. In terms of reliability, the simulations have confirmed the expectations that 16 nm be more sensitive to self-heating effects. The different assessment methodologies adopted by the two foundries did not allow a comparison of the physical capabilities of the two processes, but have shed some light on their individual margins.
These results were validated by a fully differential E-band PA prototype in each of the two technologies, which showed in both cases very good model-hardware correlation. The 16 nm FinFET prototype showed also state-of-the-art performance compared to other published works in the E-band in FinFET technologies, which further enhances the value of the proposed design methodology.
In conclusion the main contribution of this thesis is to provide an original perspective on several known technological questions and design challenges of millimeter-wave front-end circuits in CMOS technologies, proposing original solutions for each of them. This work offers multiple promising hints for future research endeavors, such as the adaptation of the layout optimization methodology to the design of transformers-based baluns, the assessment of the impact of the metal stack profile at lower operating frequencies and the application of the algorthmic design methodology to more advanced CMOS nodes.

A Appendix

A.1 Transformer Maximum Efficiency

Assuming a complex load impedance consisting of a resistance R_L with a series capacitance C_L for the low-frequency model of the transformer in Figure 5.4b, the conditions under which (5.7) is valid are:

$$\frac{1}{\omega C_L} = \omega L_s \tag{A.1}$$

$$\omega L_p = \frac{R_L}{n^2\sqrt{\frac{1}{Q_s^2} + \frac{Q_p}{Q_s}k_m^2}} \tag{A.2}$$

where $n^2 = L_s/L_p$.

Bibliography

[A+05] Semiconductor Industry Association et al. International technology roadmap for semiconductors 2005. *http://public. itrs. net*, 2005.

[AHEK+17] Akram Al-Hourani, Robin J Evans, Sithamparanathan Kandeepan, Bill Moran, and Hamid Eltom. Stochastic geometry methods for modeling automotive radar interference. *IEEE Transactions on Intelligent Transportation Systems*, 19(2):333–344, 2017.

[AJA+14] Amir Agah, Jefy Alex Jayamon, Peter M Asbeck, Lawrence E Larson, and James F Buckwalter. Multi-drive stacked-fet power amplifiers at 90 ghz in 45 nm soi cmos. *IEEE Journal of Solid-State Circuits*, 49(5):1148–1157, 2014.

[AKE+20] Wael Abdullah Ahmad, Maciej Kucharski, Arzu Ergintav, Salah Abouzaid, Jan Wessel, Herman Jalli Ng, and Dietmar Kissinger. Multimode w-band and d-band mimo scalable radar platform. *IEEE Transactions on Microwave Theory and Techniques*, 69(1):1036–1047, 2020.

[AKRH02] Ichiro Aoki, Scott D Kee, David B Rutledge, and Ali Hajimiri. Distributed active transformer-a new power-combining and impedance-transformation technique. *IEEE Transactions on Microwave Theory and Techniques*, 50(1):316–331, 2002.

[AKS+18] Irfan Ahmed, Hedi Khammari, Adnan Shahid, Ahmed Musa, Kwang Soon Kim, Eli De Poorter, and Ingrid Moerman. A survey on hybrid beamforming techniques in 5g: Architecture and system model perspectives. *IEEE Communications Surveys & Tutorials*, 20(4):3060–3097, 2018.

[ALK+08] Kyu Hwan An, Ockgoo Lee, Hyungwook Kim, Dong Ho Lee, Jeonghu Han, Ki Seok Yang, Younsuk Kim, Jae Joon Chang, Wangmyong Woo, Chang-Ho Lee, et al. Power-combining

transformer techniques for fully-integrated cmos power amplifiers. *IEEE Journal of Solid-State Circuits*, 43(5):1064–1075, 2008.

[AMB+11] Hiroki Asada, Kota Matsushita, Keigo Bunsen, Kenichi Okada, and Akira Matsuzawa. A 60ghz cmos power amplifier using capacitive cross-coupling neutralization with 16% pae. In *2011 41st European Microwave Conference*, pages 1115–1118. IEEE, 2011.

[ARCRVR+18] Ana Belén Amado-Rey, Yolanda Campos-Roca, Friedbert Van Raay, Christian Friesicke, Sandrine Wagner, Hermann Massler, Arnulf Leuther, and Oliver Ambacher. Analysis and development of submillimeter-wave stacked-fet power amplifier mmics in 35-nm mhemt technology. *IEEE Transactions on Terahertz Science and Technology*, 8(3):357–364, 2018.

[ARÖ+19] Peter M Asbeck, Narek Rostomyan, Mustafa Özen, Bagher Rabet, and Jefy A Jayamon. Power amplifiers for mm-wave 5g applications: Technology comparisons and cmos-soi demonstration circuits. *IEEE Transactions on Microwave Theory and Techniques*, 67(7):3099–3109, 2019.

[B+05] Graham M Brooker et al. Understanding millimetre wave fmcw radars. In *1st international Conference on Sensing Technology*, volume 1, 2005.

[BB+10] Mudit Ratana Bhalla, Anand Vardhan Bhalla, et al. Generations of mobile wireless technology: A survey. *International Journal of Computer Applications*, 5(4):26–32, 2010.

[Bev20] Andrea Bevilacqua. Fundamentals of integrated transformers: From principles to applications. *IEEE Solid-State Circuits Magazine*, 12(4):86–100, 2020.

[BK14] Manoj Barnela and Dr Suresh Kumar. Digital modulation schemes employed in wireless communication: A literature review. *International Journal of Wired and Wireless Communications*, 2(2):15–21, 2014.

[Bla69] James R Black. Electromigration—a brief survey and some recent results. *IEEE Transactions on Electron Devices*, 16(4):338–347, 1969.

[Boe10] Michael Boers. A 60ghz transformer coupled amplifier in 65nm digital cmos. In *2010 IEEE Radio Frequency Integrated Circuits Symposium*, pages 343–346. IEEE, 2010.

[BSRP06] Tonio Biondi, Angelo Scuderi, Egidio Ragonese, and Giuseppe Palmisano. Analysis and modeling of layout scaling in silicon integrated stacked transformers. *IEEE Transactions on Microwave Theory and Techniques*, 54(5):2203–2210, 2006.

[BY19] Rayan Bajwa and Murat Kaya Yapici. Integrated on-chip transformers: recent progress in the design, layout, modeling and fabrication. *Sensors*, 19(16):3535, 2019.

[BYC+03] Kamel Benaissa, Jau-Yuann Yang, Darius Crenshaw, Byron Williams, Seetharaman Sridhar, Johnny Ai, Gianluca Boselli, Song Zhao, Shaoping Tang, Stanton Ashburn, et al. Rf cmos on high-resistivity substrates for system-on-chip applications. *IEEE Transactions on Electron Devices*, 50(3):567–576, 2003.

[CCE20] Mengqi Cui, Corrado Carta, and Frank Ellinger. A 21-dbm 3.7 w/mm^2 28.7% pae 64-ghz power amplifier in 22-nm fd-soi. *IEEE Solid-State Circuits Letters*, 3:386–389, 2020.

[CCW+21] Kun-Da Chu, Steven Callender, Yanjie Wang, Jacques Christophe Rudell, Stefano Pellerano, and Christopher Hull. A reconfigurable non-uniform power-combining v-band pa with+ 17.9 dbm psat and 26.5% pae in 16-nm finfet cmos. *IEEE Journal of Solid-State Circuits*, 2021.

[CFH+10] Chan-Byoung Chae, Antonio Forenza, Robert W Heath, Matthew R McKay, and Iain B Collings. Adaptive mimo transmission techniques for broadband wireless communication systems [topics in wireless communications]. *IEEE Communications Magazine*, 48(5):112–118, 2010.

[CHA+13] Jing-Hwa Chen, Sultan R Helmi, Reza Azadegan, Farshid Aryanfar, and Saeed Mohammadi. A broadband stacked power amplifier in 45-nm cmos soi technology. *IEEE Journal of Solid-State Circuits*, 48(11):2775–2784, 2013.

[CHW+16] Cheng-Feng Chou, Yuan-Hung Hsiao, Yi-Ching Wu, Yu-Hsuan Lin, Chen-Wei Wu, and Huei Wang. Design of a *v*-band

20-dbm wideband power amplifier using transformer-based radial power combining in 90-nm cmos. *IEEE Transactions on Microwave Theory and Techniques*, 64(12):4545–4560, 2016.

[CK14] Anandaroop Chakrabarti and Harish Krishnaswamy. High-power high-efficiency class-e-like stacked mmwave pas in soi and bulk cmos: Theory and implementation. *IEEE Transactions on Microwave Theory and Techniques*, 62(8):1686–1704, 2014.

[CM01] Yuhua Cheng and Mishel Matloubian. High frequency characterization of gate resistance in rf mosfets. *IEEE Electron Device Letters*, 22(2):98–100, 2001.

[CNP+19] Andrea Cavarra, Claudio Nocera, Giuseppe Papotto, Egidio Ragonese, and Giuseppe Palmisano. Transformer design for 77-ghz down-converter in 28-nm fd-soi cmos technology. In *Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2018 6*, pages 195–201. Springer, 2019.

[CPH18] Steven Callender, Stefano Pellerano, and Christopher Hull. A compact 75ghz pa with 26.3% pae and 24ghz bandwidth in 22nm finfet cmos. In *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pages 224–227. IEEE, 2018.

[CPH19] Steven Callender, Stefano Pellerano, and Christopher Hull. An *e*-band power amplifier with 26.3% pae and 24-ghz bandwidth in 22-nm finfet cmos. *IEEE Journal of Solid-State Circuits*, 54(5):1266–1273, 2019.

[CQP+20] Vittorio Camarchia, Roberto Quaglia, Anna Piacibello, Duy P Nguyen, Hua Wang, and Anh-Vu Pham. A review of technologies and design techniques of millimeter-wave power amplifiers. *IEEE Transactions on Microwave Theory and Techniques*, 68(7):2957–2983, 2020.

[Cri06] Steve C Cripps. *RF power amplifiers for wireless communications*, volume 2. Artech house Norwood, MA, 2006.

[CRN09] Debopriyo Chowdhury, Patrick Reynaert, and Ali M Niknejad. Design considerations for 60 ghz transformer-coupled cmos power amplifiers. *IEEE Journal of Solid-State Circuits*, 44(10):2733–2744, 2009.

[CS17] Chandrakanth Reddy Chappidi and Kaushik Sengupta. Globally optimal matching networks with lossy passives and efficiency bounds. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(1):257–269, 2017.

[CTC+14] Xuesong Chen, Mu Kai Tsai, Chih-Hung Chen, Ryan Lee, and David C Chen. Extraction of gate resistance in sub-100-nm mosfets with statistical verification. *IEEE Transactions on Electron Devices*, 61(9):3111–3117, 2014.

[CX05] Kyuchul Chong and Ya-Hong Xie. High-performance on-chip transformers. *IEEE Electron Device Letters*, 26(8):557–559, 2005.

[DAB03] RMC De Almeida and Israel Jacob Rabin Baumvol. Reaction–diffusion in high-k dielectrics on si. *Surface Science Reports*, 49(1-3):1–114, 2003.

[DDT+20] Saeid Daneshgar, Kaushik Dasgupta, Chintan Thakkar, Anandaroop Chakrabarti, Cooper S Levy, James E Jaussi, and Bryan Casper. High-power generation for mm-wave 5g power amplifiers in deep submicrometer planar and finfet bulk cmos. *IEEE Transactions on Microwave Theory and Techniques*, 68(6):2041–2056, 2020.

[DHG+13] Hayg-Taniel Dabag, Bassel Hanafi, Fatih Golcuk, Amir Agah, James F Buckwalter, and Peter M Asbeck. Analysis and design of stacked-fet millimeter-wave power amplifiers. *IEEE Transactions on Microwave Theory and Techniques*, 61(4):1543–1556, 2013.

[DN10] Zhiming Deng and Ali M Niknejad. A layout-based optimal neutralization technique for mm-wave differential amplifiers. In *2010 IEEE Radio Frequency Integrated Circuits Symposium*, pages 355–358. IEEE, 2010.

[DOA+17] Zhou Du, Eckhard Ohlmer, Kimmo Aronkytö, Jyri Putkonen, Jouko Kapanen, and Daniel Swist. 5g e-band backhaul system measurements in urban street-level scenarios. In *2017 47th European Microwave Conference (EuMC)*, pages 372–375. IEEE, 2017.

[DR14] Noël Deferm and Patrick Reynaert. Differential and common mode stability analysis of differential mm-wave cmos ampli-

fiers with capacitive neutralization. *Analog Integrated Circuits and Signal Processing*, 80:1–12, 2014.

[DSC+12] Benjamin Dormieu, Patrick Scheer, Clément Charbuillet, Hervé Jaouen, and François Danneville. Revisited rf compact model of gate resistance suitable for high-k/metal gate technology. *IEEE Transactions on Electron Devices*, 60(1):13–19, 2012.

[EC00] Christian Enz and Yuhua Cheng. Mos transistor modeling for rf ic design. *IEEE Journal of Solid-State Circuits*, 35(2):186–201, 2000.

[EGKB07] Ouail El-Gharniti, Eric Kerherve, and Jean-Baptiste Begueret. Modeling and characterization of on-chip transformers for silicon rfic. *IEEE Transactions on Microwave Theory and Techniques*, 55(4):607–615, 2007.

[FLWL03] W Fan, Albert Lu, LL Wai, and BK Lok. Mixed-mode s-parameter characterization of differential structures. In *Proceedings of the 5th Electronics Packaging Technology Conference (EPTC 2003)*, pages 533–537. IEEE, 2003.

[Fra11] Martin M Frank. High-k/metal gate innovations enabling continued cmos scaling. In *2011 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pages 25–33. IEEE, 2011.

[GH12] Fadhel M Ghannouchi and Mohammad S Hashmi. *Load-pull techniques with applications to power amplifier design*, volume 32. Springer Science & Business Media, 2012.

[GJLY06] Wei Gao, Chao Jiao, Tao Liu, and Zhiping Yu. Scalable compact circuit model for differential spiral transformers in cmos rfics. *IEEE Transactions on Electron Devices*, 53(9):2187–2194, 2006.

[GP17] Giovanni Ghione and Marco Pirola. *Microwave Electronics*. Cambridge University Press, 2017.

[GSD+17] Davide Guermandi, Qixian Shi, Andy Dewilde, Veerle Derudder, Ubaid Ahmad, Annachiara Spagnolo, Ilja Ocket, André Bourdoux, Piet Wambacq, Jan Craninckx, et al. A 79-ghz 2×2 mimo pmcw radar soc in 28-nm cmos. *IEEE Journal of Solid-State Circuits*, 52(10):2613–2626, 2017.

[GTR+07] Kavita Goverdhanam, Youri Tretiakov, G Ali Rezvani, Sharad Kapur, and David E Long. De-embedding considerations for high q rfic inductors. In *2007 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pages 449–452. IEEE, 2007.

[GURP15] Zach Griffith, Miguel Urteaga, Petra Rowell, and Richard Pierson. 340-440mw broadband, high-efficiency e-band pa's in inp hbt. In *2015 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, pages 1–4. IEEE, 2015.

[HCM16] Sultan R Helmi, Jing-Hwa Chen, and Saeed Mohammadi. High-efficiency microwave and mm-wave stacked cell cmos soi power amplifiers. *IEEE Transactions on Microwave Theory and Techniques*, 64(7):2025–2038, 2016.

[HGPR+16] Robert W Heath, Nuria Gonzalez-Prelcic, Sundeep Rangan, Wonil Roh, and Akbar M Sayeed. An overview of signal processing techniques for millimeter wave mimo systems. *IEEE journal of selected topics in signal processing*, 10(3):436–453, 2016.

[HHG+12] Debin Hou, Wei Hong, Wang Ling Goh, Yong Zhong Xiong, Muthukumaraswamy Annamalai Arasu, Jin He, Jixin Chen, and Mohammad Madihian. Distributed modeling of six-port transformer for millimeter-wave sige bicmos circuits design. *IEEE transactions on microwave theory and techniques*, 60(12):3728–3738, 2012.

[HKGZ10] Allan Huynh, Magnus Karlsson, Shaofang Gong, and Vitaliy Zhurbenko. Mixed-mode s-parameters and conversion techniques. In *Advanced Microwave Circuits and Systems*, number 1. IntechOpen, 2010.

[HN19] Gernot Hueber and Ali M Niknejad. *Millimeter-wave Circuits for 5G and Radar*. Cambridge University Press, 2019.

[HTC08] Heng-Ming Hsu, Chien-Wen Tseng, and Kai-Yuen Chan. Characterization of on-chip transformer using microwave technique. *IEEE transactions on electron devices*, 55(3):833–837, 2008.

[Hu84] Genda J Hu. A better understanding of cmos latch-up. *IEEE Transactions on Electron Devices*, 31(1):62–67, 1984.

[JBA16] Jefy A Jayamon, James F Buckwalter, and Peter M Asbeck. A pmos mm-wave power amplifier at 77 ghz with 90 mw output power and 24% efficiency. In *2016 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pages 262–265. IEEE, 2016.

[JCV+09] Malgorzata Jurczak, Nadine Collaert, Anabela Veloso, T Hoffmann, and Serge Biesemans. Review of finfet technology. In *2009 IEEE international SOI conference*, pages 1–4. IEEE, 2009.

[JF13] Ted Johansson and Jonas Fritzin. A review of watt-level cmos rf power amplifiers. *IEEE transactions on microwave theory and techniques*, 62(1):111–124, 2013.

[JJO+18] Alvin Joseph, Vibhor Jain, Shih Ni Ong, Randy Wolf, Suh Fei Lim, and Jagar Singh. Technology positioning for mm wave applications: 130/90nm sige bicmos vs. 28nm rfcmos. In *2018 IEEE BiCMOS and compound semiconductor integrated circuits and technology symposium (BCICTS)*, pages 18–21. IEEE, 2018.

[JOC+98] Xiaodong Jin, Jia-Jiunn Ou, Chih-Hung Chen, Weidong Liu, M Jamal Deen, Paul R Gray, and Chenming Hu. An effective gate resistance model for cmos rf and noise modeling. In *International Electron Devices Meeting 1998. Technical Digest (Cat. No. 98CH36217)*, pages 961–964. IEEE, 1998.

[KBL14] Muhammad Imran Khan, Abdul Rehman Buzdar, and Fujiang Lin. Self-heating and reliability issues in finfet and 3d ics. In *2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, pages 1–3. IEEE, 2014.

[KCL12] Ju-Young Kim, Min-Kwon Choi, and Seonghearn Lee. A "thru-short-open" de-embedding method for accurate on-wafer rf measurements of nano-scale mosfets. *Journal of Semiconductor Technology and Science*, 12(1):53–58, 2012.

[KGV91] MCAM Koolen, JAM Geelen, and MPJG Versleijen. An improved de-embedding technique for on-wafer high-frequency characterization. In *Proc. Bipolar Circuits Technol. Meeting*, pages 188–191. Minneapolis, MN, 1991.

[KK15] Youngmin Kim and Youngwoo Kwon. Analysis and design of millimeter-wave power amplifier using stacked-fet structure. *IEEE Transactions on Microwave Theory and Techniques*, 63(2):691–702, 2015.

[KKJS07] Myounggon Kang, In Man Kang, Young Ho Jung, and Hyungcheol Shin. Separate extraction of gate resistance components in rf mosfets. *IEEE transactions on electron devices*, 54(6):1459–1463, 2007.

[KKS05] Myounggon Kang, In Man Kang, and Hyungcheol Shin. Extraction and modeling of gate electrode resistance in rf mosfets. In *2005 International Conference on Integrated Circuit Design and Technology, 2005. ICICDT 2005.*, pages 207–210. IEEE, 2005.

[KLM+20] Oltjon Kodheli, Eva Lagunas, Nicola Maturo, Shree Krishna Sharma, Bhavani Shankar, Jesus Fabian Mendoza Montoya, Juan Carlos Merlano Duncan, Danilo Spano, Symeon Chatzinotas, Steven Kisseleff, et al. Satellite communications in the new space era: A survey and future challenges. *IEEE Communications Surveys & Tutorials*, 23(1):70–109, 2020.

[KN14] Oleg Kononchuk and B-Y Nguyen. *Silicon-on-insulator (soi) technology: Manufacture and applications*. Elsevier, 2014.

[Kuh12] Kelin J Kuhn. Considerations for ultimate cmos scaling. *IEEE transactions on Electron Devices*, 59(7):1813–1828, 2012.

[LCR+20] Hyung-Jin Lee, Steven Callender, Said Rami, Woorim Shin, Qiang Yu, and Jose Mauricio Marulanda. Intel 22nm low-power finfet (22ffl) process technology for 5g and beyond. In *2020 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–7. IEEE, 2020.

[LGC+97] Wi Liu, R Gharpurey, MC Chang, U Erdogan, R Aggarwal, and JP Mattia. Rf mosfet modeling accounting for distributed substrate and channel resistances with emphasis on the bsim3v3 spice model. In *International Electron Devices Meeting. IEDM Technical Digest*, pages 309–312. IEEE, 1997.

[Lit01] Andrei Litwin. Overlooked interfacial silicide-polysilicon gate resistance in mos transistors. *IEEE Transactions on Electron Devices*, 48(9):2179–2181, 2001.

[LJW+07] Hongmei Li, Basanth Jagannathan, Jing Wang, Tai-Chi Su, Susan Sweeney, John J Pekarik, Yun Shi, David Greenberg, Zhenrong Jin, Robert Groves, et al. Technology scaling and device design for 350 ghz rf performance in a 45nm bulk cmos process. In *2007 IEEE Symposium on VLSI Technology*, pages 56–57. IEEE, 2007.

[LKB15] Bernardo Leite, Eric Kerhervé, and Didier Belot. Design of 28 nm cmos integrated transformers for a 60 ghz power amplifier. In *Proceedings of the 28th Symposium on Integrated Circuits and Systems Design*, pages 1–6, 2015.

[LKBB09] Bernardo Leite, Eric Kerhervé, Jean-Baptiste Bégueret, and Didier Belot. Transformer topologies for mmw integrated circuits. In *2009 European Microwave Conference (EuMC)*, pages 181–184. IEEE, 2009.

[LKBB12] Bernardo Leite, Eric Kerherve, Jean-Baptiste Bégueret, and Didier Belot. An analytical broadband model for millimeter-wave transformers in silicon technologies. *IEEE Transactions on electron devices*, 59(3):582–589, 2012.

[Lou16] Errikos Lourandakis. *On-wafer Microwave Measurements and De-embedding*. Artech House, 2016.

[LSC+20] Zheng Liu, Tushar Sharma, Chandrakanth Reddy Chappidi, Suresh Venkatesh, Yiming Yu, and Kaushik Sengupta. A 42-62 ghz transformer-based broadband mm-wave inp pa with second-harmonic waveform engineering and enhanced linearity. *IEEE Transactions on Microwave Theory and Techniques*, 2020.

[MNQ+99] Sebastien Mons, J-C Nallatamby, Raymond Quéré, P Savary, and Juan Obregon. A unified approach for the linear and nonlinear stability analysis of microwave circuits using commercially available tools. *IEEE Transactions on Microwave Theory and Techniques*, 47(12):2403–2409, 1999.

[MP05] Michael Marcus and Bruno Pattan. Millimeter wave propagation: spectrum management implications. *IEEE Microwave Magazine*, 6(2):54–62, 2005.

[MPR+15] Kenichi Miyaguchi, Bertrand Parvais, Lars-Åke Ragnarsson, Piet Wambacq, Praveen Raghavan, Abdelkarim Mercha, Anda

Mocuta, Diederik Verkest, and Aaron Thean. Modeling finfet metal gate stack resistance for 14nm node and beyond. In *2015 International Conference on IC Design & Technology (ICICDT)*, pages 1–4. IEEE, 2015.

[MSM15] Arash Maskooki, Gabriele Sabatino, and Nathalie Mitton. Analysis and performance evaluation of the next generation wireless networks. In *Modeling and Simulation of Computer Networks and Systems*, pages 601–627. Elsevier, 2015.

[NCC12] Ali M Niknejad, Debopriyo Chowdhury, and Jiashu Chen. Design of cmos power amplifiers. *IEEE Transactions on Microwave Theory and Techniques*, 60(6):1784–1796, 2012.

[NRB02] Kiat T Ng, Behzad Rejaei, and Joachim N Burghartz. Substrate effects in monolithic rf transformers on silicon. *IEEE Transactions on Microwave Theory and Techniques*, 50(1):377–383, 2002.

[NRS+22] Lucas Nyssens, Martin Rack, Christoph Schwan, Zhixing Zhao, Steffen Lehmann, Tom Hermann, Frederic Allibert, Cécile Aulnette, Dimitri Lederer, and Jean-Pierre Raskin. Impact of substrate resistivity on spiral inductors at mm-wave frequencies. *Solid-State Electronics*, 194:108377, 2022.

[NW20] Huy Thong Nguyen and Hua Wang. A coupler-based differential mm-wave doherty power amplifier with impedance inverting and scaling baluns. *IEEE Journal of Solid-State Circuits*, 55(5):1212–1223, 2020.

[OLC+18] SN Ong, S Lehmann, WH Chow, C Zhang, C Schippel, LHK Chan, Y Andee, M Hauschildt, KKS Tan, J Watts, et al. A 22nm fdsoi technology optimized for rf/mmwave applications. In *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pages 72–75. IEEE, 2018.

[PAG+14] John J Pekarik, J Adkisson, P Gray, Q Liu, R Camillo-Castillo, M Khater, V Jain, B Zetterlund, A DiVergilio, X Tian, et al. A 90nm sige bicmos technology for mm-wave and high-performance analog applications. In *2014 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, pages 92–95. IEEE, 2014.

[PAO+02] SJ Pearton, CR Abernathy, ME Overberg, GT Thaler, AH Onstine, BP Gila, F Ren, B Lou, and J Kim. New applications advisable for gallium nitride. *Materials today*, 5(6):24–31, 2002.

[Poz11] David M Pozar. *Microwave engineering*. John wiley & sons, 2011.

[PPS+13] Sangyoung Park, Jaehyun Park, Donghwa Shin, Yanzhi Wang, Qing Xie, Massoud Pedram, and Naehyuck Chang. Accurate modeling of the delay and energy overhead of dynamic voltage and frequency scaling in modern microprocessors. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32(5):695–708, 2013.

[RK05] Marco Racanelli and Paul Kempf. Sige bicmos technology for rf circuit applications. *IEEE transactions on electron devices*, 52(7):1259–1270, 2005.

[RKS17] Usman Raza, Parag Kulkarni, and Mahesh Sooriyabandara. Low power wide area networks: An overview. *ieee communications surveys & tutorials*, 19(2):855–873, 2017.

[RNC+20] Egidio Ragonese, Claudio Nocera, Andrea Cavarra, Giuseppe Papotto, Simone Spataro, and Giuseppe Palmisano. A comparative analysis between standard and mm-wave optimized beol in a nanoscale cmos technology. *Electronics*, 9(12):2124, 2020.

[RNN+22] Martin Rack, Lucas Nyssens, Massinissa Nabet, C Schwan, Z Zhao, S Lehmann, T Herrmann, D Henke, A Kondrat, C Soonekindt, et al. High-resistivity substrates with pn interface passivation in 22 nm fd-soi. In *2022 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, pages 1–2. IEEE, 2022.

[RS06] Patrick Reynaert and Michiel Steyaert. *RF power amplifiers for mobile communications*. Springer Science & Business Media, 2006.

[RW15] John Robertson and Robert M Wallace. High-k materials and metal gates for cmos applications. *Materials Science and Engineering: R: Reports*, 88:1–41, 2015.

[RYL94] Behzad Razavi, Ran-Hong Yan, and Kwing F Lee. Impact of distributed gate resistance on the performance of mos devices. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 41(11):750–754, 1994.

[RZN19] Ali Razavieh, Peter Zeitzoff, and Edward J Nowak. Challenges and limitations of cmos scaling for finfet and beyond architectures. *IEEE Transactions on Nanotechnology*, 18:999–1004, 2019.

[Sch05] Martin Schneider. Automotive radar-status and trends. In *German microwave conference*, pages 144–147. Citeseer, 2005.

[SCS+18] Peter Sagazio, Steven Callender, Woorim Shin, Oner Orhan, Stefano Pellerano, and Christopher Hull. Architecture and circuit choices for 5g millimeter-wave beamforming transceivers. *IEEE communications magazine*, 56(12):186–192, 2018.

[Sha49] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[SHBR05] Hisao Shigematsu, Tatsuya Hirose, Forrest Brewer, and Mark Rodwell. Millimeter-wave cmos circuit design. *IEEE Transactions on Microwave Theory and Techniques*, 53(2):472–477, 2005.

[SJR06] Almudena Suárez, Sanggeun Jeon, and David Rutledge. Stability analysis and stabilization of power amplifiers. *IEEE microwave magazine*, 7(5):51–65, 2006.

[Sko80] Merrill Ivan Skolnik. Introduction to radar systems. *New York*, 1980.

[SPD+16] Sherif Shakib, Hyun-Chul Park, Jeremy Dunworth, Vladimir Aparin, and Kamran Entesari. A highly efficient and linear power amplifier for 28-ghz 5g phased array radios in 28-nm cmos. *IEEE Journal of Solid-State Circuits*, 51(12):3020–3036, 2016.

[SRT+13] Keisuke Shinohara, Dean C Regan, Yan Tang, Andrea L Corrion, David F Brown, Joel C Wong, John F Robinson, Helen H Fung, Adele Schmitz, Thomas C Oh, et al. Scaling of gan hemts and schottky diodes for submillimeter-wave mmic applications. *IEEE Transactions on Electron Devices*, 60(10):2982–2996, 2013.

[SSKJ87] Bing J Sheu, Donald L Scharfetter, P-K Ko, and M-C Jeng. Bsim: Berkeley short-channel igfet model for mos transistors. *IEEE Journal of Solid-State Circuits*, 22(4):558–566, 1987.

[Sua15] Almudena Suarez. Check the stability: Stability analysis methods for microwave circuits. *IEEE Microwave Magazine*, 16(5):69–90, 2015.

[TF03] Vishal P Trivedi and Jerry G Fossum. Scaling fully depleted soi cmos. *IEEE Transactions on Electron devices*, 50(10):2095–2103, 2003.

[THJB05] Luuk F Tiemeijer, Ramon J Havens, André BM Jansman, and Yann Bouttement. Comparison of the" pad-open-short" and" open-short-load" deembedding techniques for accurate on-wafer rf characterization of high-quality passives. *IEEE Transactions on Microwave Theory and Techniques*, 53(2):723–729, 2005.

[TNH14] Siva V Thyagarajan, Ali M Niknejad, and Christopher D Hull. A 60 ghz drain-source neutralized wideband linear power amplifier in 28 nm cmos. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 61(8):2253–2262, 2014.

[TNM+20] Xinyan Tang, Johan Nguyen, Alaaeldien Medra, Khaled Khalaf, Akshay Visweswaran, Björn Debaillie, and Piet Wambacq. Design of d-band transformer-based gain-boosting class-ab power amplifiers in silicon technologies. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67(5):1447–1458, 2020.

[TP19] Van-Son Trinh and Jung-Dong Park. Theory and design of impedance matching network utilizing a lossy on-chip transformer. *IEEE Access*, 7:140980–140989, 2019.

[TP21] Van-Son Trinh and Jung-Dong Park. An 85-ghz power amplifier utilizing a transformer-based power combiner operating beyond the self-resonance frequency. *IEEE Journal of Solid-State Circuits*, 57(3):882–891, 2021.

[TPAG13] Luuk F Tiemeijer, Ralf MT Pijper, Cristian Andrei, and Emmanuel Grenados. Analysis, design, modeling, and characterization of low-loss scalable on-chip transformers. *IEEE trans-*

actions on microwave theory and techniques, 61(7):2545–2557, 2013.

[TRML+12] Julio C Tinoco, Silvestre Salas Rodriguez, Andrea G Martinez-Lopez, Joaquín Alvarado, and Jean-Pierre Raskin. Impact of extrinsic capacitances on finfet rf performance. *IEEE transactions on microwave theory and techniques*, 61(2):833–840, 2012.

[UPR+11] M Urteaga, Richard Pierson, Petra Rowell, Vibhor Jain, Evan Lobisser, and Mark JW Rodwell. 130nm inp dhbts with ft> 0.52 thz and f max> 1.1 thz. In *69th Device Research Conference*, pages 281–282. IEEE, 2011.

[VR17] Marco Vigilante and Patrick Reynaert. A wideband class-ab power amplifier with 29–57-ghz am–pm compensation in 0.9-v 28-nm bulk cmos. *IEEE Journal of Solid-State Circuits*, 53(5):1288–1301, 2017.

[VRC+11] Anabela Veloso, L-Å Ragnarsson, Moon Ju Cho, Katia Devriendt, Kristof Kellens, Farid Sebaai, Samuel Suhard, Stephan Brus, Yvo Crabbe, Tom Schram, et al. Gate-last vs. gate-first technology for aggressively scaled eot logic/rf cmos. In *2011 Symposium on VLSI Technology-Digest of Technical Papers*, pages 34–35. IEEE, 2011.

[WC05] Wen Wu and Mansun Chan. Gate resistance modeling of multifin mos devices. *IEEE electron device letters*, 27(1):68–70, 2005.

[WCC19] Zhongmin Wang, Tianying Chang, and Hong-Liang Cui. Review of active millimeter wave imaging techniques for personnel security screening. *IEEE Access*, 7:148336–148350, 2019.

[WLC+13] Shien-Yang Wu, Colin Yu Lin, MC Chiang, JJ Liaw, JY Cheng, SH Yang, Ming Liang, Tadakazu Miyashita, CH Tsai, BC Hsu, et al. A 16nm finfet cmos technology for mobile soc and computing applications. In *2013 IEEE International Electron Devices Meeting*, pages 9–1. IEEE, 2013.

[WR19] Eric Wagner and Gabriel M Rebeiz. Single and power-combined linear e-band power amplifiers in 0.12-µm sige with 19-dbm average power 1-gbaud 64-qam modulated waveforms.

IEEE Transactions on Microwave Theory and Techniques, 67(4):1531–1543, 2019.

[WS15] Hua Wang and Kaushik Sengupta. *RF and mm-Wave Power Generation in Silicon*. Academic Press, 2015.

[WW21] Fei Wang and Hua Wang. A broadband linear ultra-compact mm-wave power amplifier with distributed-balun output network: Analysis and design. *IEEE Journal of Solid-State Circuits*, 56(8):2308–2323, 2021.

[WWN+20] Hua Wang, Fei Wang, Huy Thong Nguyen, Sensen Li, Tzu-Yuan Huang, Amr S Ahmed, Michael Edward Duffy Smith, Naga Sasikanth Mannem, and Jeongseok Lee. Power amplifiers performance survey 2000-present. *PA_survey. html*, 2020.

[YGH+22] Qiang Yu, Jeffrey Garrett, Seahee Hwangbo, Georgios Dogiamis, and Said Rami. An f-band power amplifier with skip-layer via achieving 23.8% pae in finfet technology. In *2022 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pages 179–182. IEEE, 2022.

[YGT+07] Terry Yao, Michael Q Gordon, Keith KW Tang, Kenneth HK Yau, Ming-Ta Yang, Peter Schvan, and Sorin P Voinigescu. Algorithmic design of cmos lnas and pas for 60-ghz radio. *IEEE Journal of Solid-State Circuits*, 42(5):1044–1057, 2007.

[YMYZ15] Wanxin Ye, Kaixue Ma, Kiat Seng Yeo, and Qiong Zou. A 65 nm cmos power amplifier with peak pae above 18.9% from 57 to 66 ghz using synthesized transformer-based matching network. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 62(10):2533–2543, 2015.

[YXXL19] Ping Yang, Yue Xiao, Ming Xiao, and Shaoqian Li. 6g wireless communications: Vision and potential techniques. *IEEE network*, 33(4):70–75, 2019.

[ZAA21] Yu Zhang, Muhammad Alrabeiah, and Ahmed Alkhateeb. Reinforcement learning of beam codebooks in millimeter wave and terahertz mimo systems. *IEEE Transactions on Communications*, 70(2):904–919, 2021.

[ZCT+16] Xi Zhang, Daniel Connelly, Hideki Takeuchi, Marek Hytha, Robert J Mears, and Tsu-Jae King Liu. Comparison of soi

versus bulk finfet technologies for 6t-sram voltage scaling at the 7-/8-nm node. *IEEE Transactions on Electron Devices*, 64(1):329–332, 2016.

[ZGZ+21] Dixian Zhao, Peng Gu, Jiecheng Zhong, Na Peng, Mengru Yang, Yongran Yi, Jiajun Zhang, Pingyang He, Yuan Chai, Zhihui Chen, et al. Millimeter-wave integrated phased arrays. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(10):3977–3990, 2021.

[ZHW22] Xicheng Zhu, Tongde Huang, and Wen Wu. A wideband 77-ghz power amplifier with mixed matching network in 130-nm bicmos technology. In *2022 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, pages 159–161. IEEE, 2022.

[ZLO+21] Zhixing Zhao, Steffen Lehmann, Wei Lun Oo, Amit Kumar Sahoo, Shafi Syed, Quang Huy Le, Dang Khoa Huynh, Talha Chohan, Dirk Utess, Dominik Kleimaier, et al. 22fdsoi device towards rf and mmwave applications. In *2021 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)*, pages 1–6. IEEE, 2021.

[ZR13] Dixian Zhao and Patrick Reynaert. A 60-ghz dual-mode class ab power amplifier in 40-nm cmos. *IEEE Journal of Solid-State Circuits*, 48(10):2323–2337, 2013.

[ZR14] Dixian Zhao and Patrick Reynaert. An e-band power amplifier with broadband parallel-series power combiner in 40-nm cmos. *IEEE Transactions on Microwave Theory and Techniques*, 63(2):683–690, 2014.

[ZR15] Dixian Zhao and Patrick Reynaert. A 40-nm cmos e-band 4-way power amplifier with neutralized bootstrapped cascode amplifier and optimum passive circuits. *IEEE transactions on microwave theory and techniques*, 63(12):4083–4089, 2015.

Own Publications

[1] Carla Moran Guizan, Peter Baumgartner, Stefan Heinen, and Mario Lauritano. Millimeter-wave tunable impedance matching network in an advanced cmos process. *IEEE Transactions on Microwave Theory and Techniques*, 2024.

[2] Mario Lauritano and Peter Baumgartner. Optimal test structures for the characterization of integrated transformers at mm-wave frequencies using the open/thru de-embedding technique. In *2022 IEEE 34th International Conference on Microelectronic Test Structures (ICMTS)*, pages 1–4. IEEE, 2022. ©2022 IEEE.

[3] Mario Lauritano, Peter Baumgartner, Puneet Singh, Ahmet Çağri Ulusoy, and Carla Moran Guizan. Systematic design methodology for cmos millimeter-wave power amplifiers with an *e*-band fully differential implementation in 16-nm finfet. *IEEE Transactions on Microwave Theory and Techniques*, 2024. ©2024 IEEE.

[4] Mario Lauritano, Peter Baumgartner, and Ahmet Çağri Ulusoy. Test structures for the characterization of the gate resistance in 16 nm finfet rf transistors. *Electronics*, 12(14):3011, Jul 2023.

[5] Mario Lauritano, Peter Baumgartner, Ahmet Çağri Ulusoy, and Jasmin Aghassi-Hagmann. Matching network efficiency: the new old challenge for millimeter-wave silicon power amplifiers. *IEEE Microwave Magazine*, 22(12):86–96, 2021. ©2021 IEEE.

VITA

Mario Lauritano

2010-2013	Bachelor of Science in Electronic Engineering at Polytechnic University of Turin, Italy.
2013-2015	Master of Science in Electronic Engineering, RF Design, at Polytechnic University of Turin, Italy.
2014-2016	Master of Science in Electronic and Computer Engineering at University of Illinois at Chicago (UIC), USA, in partnership with Polytechnic University of Turin, Italy.
2015-2016	RF Engineering Internship at Knowles Electronics, Itasca, Illinois (USA).
2016-2018	Component Verification Engineer at Intel Corporation, Villach, Austria.
2018-2024	PhD Researcher at Intel Corporation, Neubiberg, Germany.
2022-2024	External PhD Student at Karlsruhe Institute of Technology (KIT), Institute of High-Frequency Techniques and Electronics (IHE), Karlsruhe, Germany.