ALMA MATER STUDIORIUM · UNIVERSITÀ DI BOLOGNA

Dipartimento di Ingegneria dell'Energia Elettrica e dell'Informazione
"Gugliemo Marconi" - DEI

Dottorato in Ingegneria Elettronica, Telecomunicazioni e Tecnologie dell'Informazione - Ciclo XXXIII

Settore Concorsuale: 09/E3 - ELETTRONICA

Settore Scientifico Disciplinare: ING-INF/01 - ELETTRONICA

MEASUREMENT TECHNIQUES FOR THE CHARACTERIZATION OF RADIO FREQUENCY GALLIUM NITRIDE DEVICES AND POWER AMPLIFIERS

Presentata da:
ALBERTO MARIA ANGELOTTI

Coordinatore Dottorato:
Prof. A. COSTANZO

Relatore:
Prof. A. SANTARELLI

Esame Finale Anno 2021

Abstract

The rapid growth of mobile telecommunications has fueled the development of the fifth generation (5G) of standards. This technology aims to achieve high data transmission rates and low latency, supporting a variety of novel applications with wide-ranging potential. These capabilities leverage on the use of new regions of spectrum in the Ka band, the adoption of wider channel bandwidths and the deployment of high spectral-efficiency modulation formats.
The effective deployment of 5G relies on the technological evolution of radio-frequency hardware with increased performance. The operation in high-power and high-frequency conditions is particularly challenging for power amplifiers (PA) in transmission stages, which should strive to minimize energy expenditure while maintaining low nonlinear distortion over large modulation bandwidths. In order to comply with these demanding specifications and improve the overall linearity-efficiency tradeoff, several novel architectural and technological improvements have been introduced in PA design.
In this respect, Gallium Nitride (GaN) is a promising semiconductor material, with exceptional performance in the realization of active devices for PAs in high-frequency and high-power applications. Despite many attractive properties, GaN high-electron mobility transistors (HEMT) and PAs have been shown to suffer from a variety of undesirable dynamical phenomena, which can reduce RF PA performance in terms of power, linearity and reliability. The severity of these degradation is particularly marked under the wideband modulated conditions encountered in modern transmitters.
Charge trapping has been shown to be a major contributor to these spurious effects in GaN. Its origin lies in the presence of physical imperfections in the GaN HEMT structure, due to lattice mismatches between different materials and either intentional or unintentional defects introduced during the epitaxial growth process. These defects dynamically capture and release mobile charges depending on the applied voltages and temperature, modulating the field and charge distribution of the device and ultimately affecting its electrical characteristics. Under these premises, this work investigates the impact of trapping on the behavior of GaN HEMTs and PAs using different techniques and measurement setups.
At low-frequency (LF), where trapping can be directly studied, the charge dynamics is analyzed using pulsed current transient characterizations. In this way, the capture and

release time constants and energy levels are identified in different state-of-the-art GaN HEMT technologies for 5G applications, highlighting a large variety of effects.
Instead, at high-frequency, several different methods are used in order to evaluate the impact of trapping dynamics on the RF operation of GaN devices. Tailored measurement setups have been devised in order to perform double-pulsed isodynamic S-parameters, pulsed RF and two-tone characterizations on GaN HEMTs, revealing a complex conversion behavior between LF and RF dynamics. Moreover, a new technique based on the best-linear-approximation framework is introduced in order to assess the impact of trapping on the distortion performance of GaN PAs under wideband modulated excitations.
As charge trapping is particularly sensitive to the instantaneous applied voltages, the characterization conditions should closely match the ones encountered in the final application. For 5G and similar applications, this involves the excitation of the device under test with broadband modulated signals, while at the same providing well-defined output loading conditions across the whole bandwidth. Therefore, an innovative wideband active load pull (WALP) setup is developed and presented in this work. Differently from existing techniques, the proposed WALP method uses just relative frequency-by-frequency acquisitions, such as those provided by regular vector-network-analyzers. In addition, the implications of performing modulation distortion characterizations under load-pull conditions are theoretically and experimentally studied.
Finally, an efficient implementation of a modified-Volterra model for GaN RF PAs with memory effects is presented. The technique makes use of a custom vector-fitting algorithm in order to simplify the nonlinear integral operators and enable their realization in circuit CAD or other simulation environments.

Acknowledgments

"The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly, that Alice had not a moment to think about stopping herself, before she found herself falling down what seemed a deep well. Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her, and to wonder what would happen next."

Lewis Carroll
"Alice's Adventures in Wonderland"

I spent most of the time during the PhD feeling like Alice tumbling down the rabbit hole: wondering how and why I got there or what would happen next. Undoubtedly, it has been a very tough and demanding journey, from day one till the end. Other times, it really felt like Wonderland. I had the privilege of working with many talented people and playing around with some very expensive instruments. And, most importantly, I had fun investigating a lot of interesting stuff.
It is probably impossible to objectively assess whether it was a mostly positive or negative experience, overall. It has surely been a challenging and memorable adventure that has taught me how to deal with all the uncertainties you encounter in research and, more in general, in life. Sorting them out is a completely different topic (and always a lot of hard work). In this spirit, I would like to personally acknowledge and thank all the people that have helped me throughout these three-and-a-half years.
First and foremost, I would like to express all my gratitude to Prof. Alberto Santarelli. His scientific guidance, motivation and patience have been fundamental at all stages of my PhD. Moreover, I would also like to thank him for the flawless handling all of the bureaucratic, financial and organizational matters, which is never a trivial task. I would also like to thank Prof. Schreurs and Prof. Pirola in the thesis review committee for the thorough analysis of this manuscript and for the critical advice they gave in order to improve it.
Special thanks go to Dr. Gian Piero Gibiino, who has been an exceptional teammate, research partner and day-to-day supervisor. His can-do attitude, dedication, hard work, and scientific advisory have been invaluable in helping me make the most out of this PhD and will surely be an inspiration for the years to come. I want also to thank all

the other researchers of the EDM-Lab Group: Prof. Fabrio Filicori, Prof. Pier Andrea Traverso, Prof. Rudi Paganelli and my office mate Dr. Maurizio Tinti. The stimulating discussions during these years greatly contributed to my personal and professional growth.
I would like to especially thank Prof. Corrado Florian, for energetically promoting a lot of very interesting research opportunities and for the great collaboration and supervision in tackling always new challenges.
A particular mention goes the MSc students that I had the pleasure to co-supervise during their thesis projects: Alessio, France, Armand and Mattia. The experience of working with each of them has been extremely rewarding for me, and I hope that they learned something useful. I am really proud that Alessio and Mattia decided to embark on a PhD at EDM-Lab and wish them all the luck for it.
I would also like to thank all the researchers that I met during my five months at KULeuven for the hosting me in an amazing research environment and for the fruitful cooperation that resulted from this.
I am especially grateful for the unique opportunity of collaborating with Keysight Technologies. Their assistance and equipment have made possible a sizable part of the research in this work. I would specifically like to express my thanks to Dr. Troels Nielsen for his time and his great research insights.
In these years, I have been truly blessed by the presence and encouragement of all my friends and family, through the good times and bad. That has made all the difference to me. To the people who have crossed my path in the Bologna 3 scout group: you will always hold a special place in my heart.
Finally, mom and dad. Your unconditional love and support have brought me here. I just hope you are proud of this achievement.

Contents

List of Figures

List of Tables

Acronyms

5G Fifth Generation Mobile Communications Standard

ACPR Adjacent Channel Power Ratio

ADC Analog-to-digital converter

Al Aluminum

AlGaN Aluminum Gallium Nitride

AlN Aluminum Nitride

AWG Arbitrary Waveform Generator

C Carbon

CW Continuous Wave

DAC Digital-to-analog converter

DHEMT Double Heterostructure High Electron Mobility Transistor

DHFET Double Heterostructure Field Effect Transistor

DPD Digital Predistortion

EVM Error Vector Magnitude

Fe Iron

FET Field Effect Transistor

GaAs Gallium Arsenide

GaN Gallium Nitride

HEMT High Electron Mobility Transistor

IM Intermodulation

IMD Intermodulation Distortion

LDMOS Laterally Diffused Metal Oxide Semiconductor

LO Local Oscillator

LP Load Pull

LSOP Large Signal Operating Point

MMIC Microwave Monolithic Integrated Circuit

NMSE Normalized Mean Square Error

NVNA Nonlinear Vector Network Analyzer

OFDM Orthogonal Frequency Division Modulation

PA Power Amplifier

PAE Power Added Efficiency

PAPR Peak to Average Power Ratio

RF Radio Frequency

Si Silicon

SiC Silicon Carbide

SNR Signal-to-noise ratio

SOLT Short Open Load Thru

VNA Vector Network Analyzer

VSA Vector Signal Analyzer

VSG Vector Signal Generator

VSWR Voltage Standing Wave Ratio

WALP Wideband Active Load Pull

Chapter 1

Introduction

1.1 RF electronics for telecommunications applications

Mobile telecommunications have experienced incredible growth in terms of volume and capabilities in the last thirty years. Their widespread use and availability has profoundly shaped the modern world, with a wide-ranging impact across all segments of society. The ever-increasing demand for access to data services from mobile users is pushing industry and research into the fifth generation (5G) of wireless telecommunications networks [1]. These new technologies and standards aim at achieving multi-Gb/s transmission rates and low latency, enabling diverse applications such as multimedia streaming, Internet of Things (IoT) and Big Data [2].
The evolution in capabilities leverages on the presence of large portions of still unlicensed spectrum at high-frequencies. In this way, ever-wider modulation bandwidths (hundreds of MHz to GHz) can be accessed to support the required performance without interfering with current low-frequency telecommunications in L and S bands. For 5G, while the FR1 frequency range still shares the sub-6-GHz band with previous standards, the new FR2 range will target several bands in the 24.25-43.5 GHz frequency interval [3]. Moreover, given that bandwidth is a costly, limited and heavily regulated resource, modern standards adopt digital modulation schemes, such as orthogonal-frequency-division mouldation (OFDM), designed to have high spectral efficiency.
At the same time, energy efficiency remains a priority, particularly in high-power components. Devices and electronic systems operating in low-efficiency conditions waste large amounts of electrical energy as heat. Moreover, effective thermal dissipation is required, as high-temperature operation can reduce component reliability and lifetime. From a network perspective, the cost of operating basestation infrastructure is largely determined by the energy consumption to power and cool the internal components. From the mobile user's side, less energy usage entails longer battery durations and greater autonomy. In

both cases, the RF power amplifier (PA) in the final stage of the transmission chain constitutes a major contributor to the overall consumption and therefore is a preferred target for improvements of the system efficiency [4].
However, efficient amplification of broadband high-frequency communications signals poses a number of challenges in the design and characterization of amplifiers and transistors to be used within wireless transmitters. Modulated signal that achieve good spectral efficiency typically employ time-domain waveforms with amplitude-distributions with elevated peak-to-average-power-ratio (PAPR), with OFDM known to display PAPR values in excess of 10 dB [5]. Therefore, the output power remains close to its average value most of the time, with infrequent high-power peaks.
This operating condition is particularly problematic: in order not to irreversibly distort the signal peaks, traditional RF-PA architectures [6] are designed in high linearity operating classes (e.g., A-B). At the same time, these classes operate at low efficiency when the input signal is in back-off conditions with respect to the peak, that is, for the great majority of time. This leads to an extremely poor average efficiency, often just a few percents, with subsequent problems due to increased heat dissipation and power consumption. The use of more efficient operating classes, while indeed possible [7], invariably leads to the generation nonlinear distortion in the amplifier, with linearity performance that may become unacceptable in terms of compliance with spectral masks and error vector magnitude specifications [3].
Therefore, the design of RF power amplifiers inevitably involves a fundamental trade-off between linearity and power efficiency, particularly in broadband operating conditions. In order to push the boundaries of this tradeoff and optimize transmitter performance, several solutions have been proposed in literature:

- *Advanced PA architectures.* Single transistor topologies, and their cascade in multiple stages, are standard in order to achieve power amplification at RF frequencies [6]. However, this type of standard solutions offer little means in improving the linearity-efficiency tradeoff. Therefore, several novel architectures have been devised, featuring specific combinations of different PA blocks in order to improve the overall efficiency, particularly in back-off conditions. Some relevant examples are Doherty [8], load-modulated balanced [9] and Chireix/outphasing amplifiers [10].

- *Digital Predistortion (DPD).* The operation of an amplifier into a non-linear working region can boost its efficiency at the price of a significant amount of output distortion. Digital predistortion (DPD) involves a measurement-based identification of a nonlinear-dynamic pre-inverse for the PA, such that the cascade of the two systems becomes equivalent to a linear amplification stage. This type of pre-compensation is performed in the digital baseband domain by directly modifying the transmitted constellation symbols. DPD is often used in combination with other solutions, in order to compensate for the additional distortion introduced by efficiency-enhancing techniques [11].

- *Supply modulation*. In envelope tracking [12], the supply voltage of the RF PA is dynamically modified using an additional baseband supply modulator amplifier, in order to follow the input envelope amplitude. In this way, the infrequent high-power peaks are amplified linearly using a large supply voltage, while the amplification of the rest of the signal is performed using a lower supply, thus saving energy. The design of the supply modulator is particularly critical [13] in order to ensure that the complete assembly will display higher efficiency than the RF PA alone. Envelope tracking can be applied to several PA architectures, but usually introduces an additional source of nonlinearity that has to be compensated using DPD [14]. The supply modulation technique can also be used as a starting point for other efficiency-enhancing architectures, such as envelope elimination and restoration [15]

In order to meet the stringent specifications imposed by the new standards and improve overall performance, these advancements in PA architectures have to be matched by the development of novel active device technologies. In this respect, solid-state semiconductor devices have replaced other solutions (e.g., traveling waves tubes) in virtually all PA applications. RF electronics makes use of several different transistor and material technologies, which can lead to superior performance in specific applications, with Gallium Nitride (GaN) and other III-V compounds being particularly suited for the realization of PAs. The key characteristics of GaN are analyzed in the following section.

1.2 GaN technology

1.2.1 Semiconductors for radio-frequency electronics

Several different semiconductor technologies are available for the realization of power amplifiers at RF and microwave frequencies. As shown in the survey reported in Fig. 1.1 [16], the use of Silicon LDMOS or CMOS processes is common in the low-frequency and low-power ranges respectively, thanks to the low cost, high integration, high performance and high reliability of mature Si-based platforms. Instead, various III-V compounds are typically employed in high-frequency and high-power applications, thanks to the favorable electrical characteristics of the employed materials. Despite having higher performance, these semiconductors have less developed technological processes and higher overall costs, limiting their use to niche applications. Gallium Nitride, in particular, has found widespread usage in recent years. While GaN popularity was initially fueled by its role in power-electronics (e.g., switching converters) [17], the material displays excellent high-frequency capabilities, with operation in the microwave and sub-mm-wave ranges that have been traditionally covered by Gallium Arsenide (GaAs).
The rise of GaN in high-power and high-frequency applications can be explained in its unique combination of physical characteristics, which are compared with the ones of other semiconductors in Table 1.1. The wide bandgap voltage of GaN enables the use of high

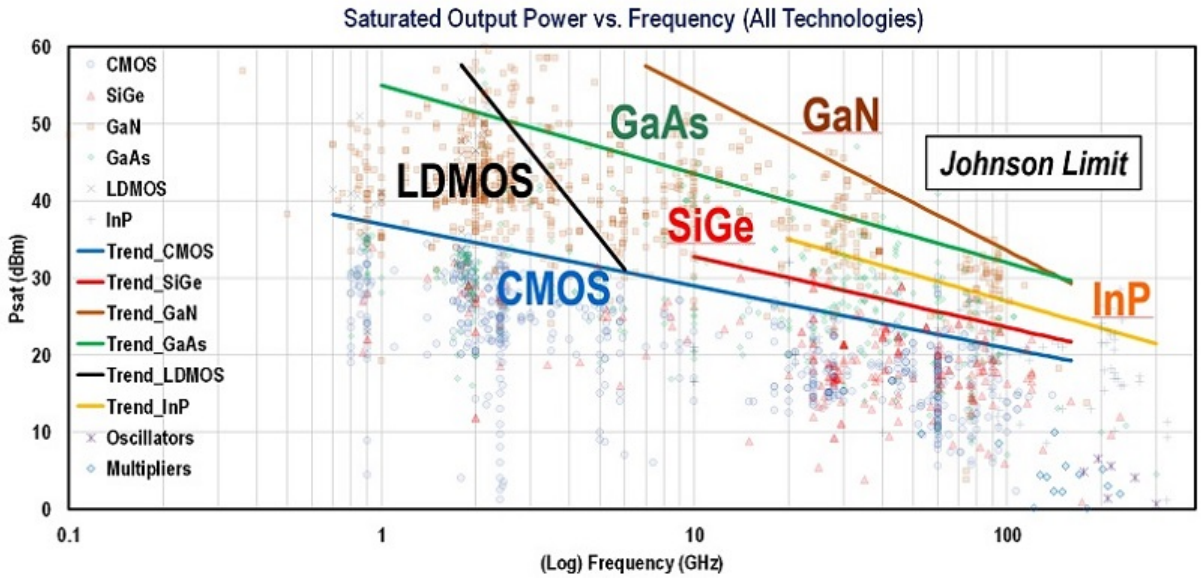


Figure 1.1: Survey of PA semiconductor technology per frequency range and saturated power [16].

Parameter	**Si**	**GaAs**	**SiC**	**GaN**
Bandgap (eV)	1.1	1.42	3.26	3.39
Breakdown electric field $\left(\frac{MV}{cm}\right)$	0.3	0.4	3.0	3.3
Dielectric Constant	11.8	13.1	10	9
Electron mobility $\left(\frac{cm^2}{Vs}\right)$	1350	8500	700	1200
Electron saturation velocity $(10^7 \frac{cm}{s})$	1	1	2	2.5
Thermal Conductivity $\left(\frac{W}{cms}\right)$	1.5	0.43	3.3-4.5	1.3
Lattice Constant (nm)	54	57	31	32
Johnson figure of merit	1	2.7	20	27.5

Table 1.1: Electrical characteristics of different RF semiconductor materials.

voltages, and therefore RF power, and the possibility of operating at higher junction temperatures with respect of other materials. The high-voltage capabilities are also enhanced by an extremely large breakdown field and better thermal conductivity than GaAs. This features allow to realize compact GaN MMICs with extremely high power densities, without having to resort to external combining of several amplifiers [18]. The possibility of higher output voltage swings, and correspondently lower current for a given output power, also greatly simplifies RF loadline matching conditions. The lower mobility with respect to GaAs, which could in principle limit dc gain and maximum operating frequency of GaN devices, is offset by the larger saturation velocity which ultimately determines the

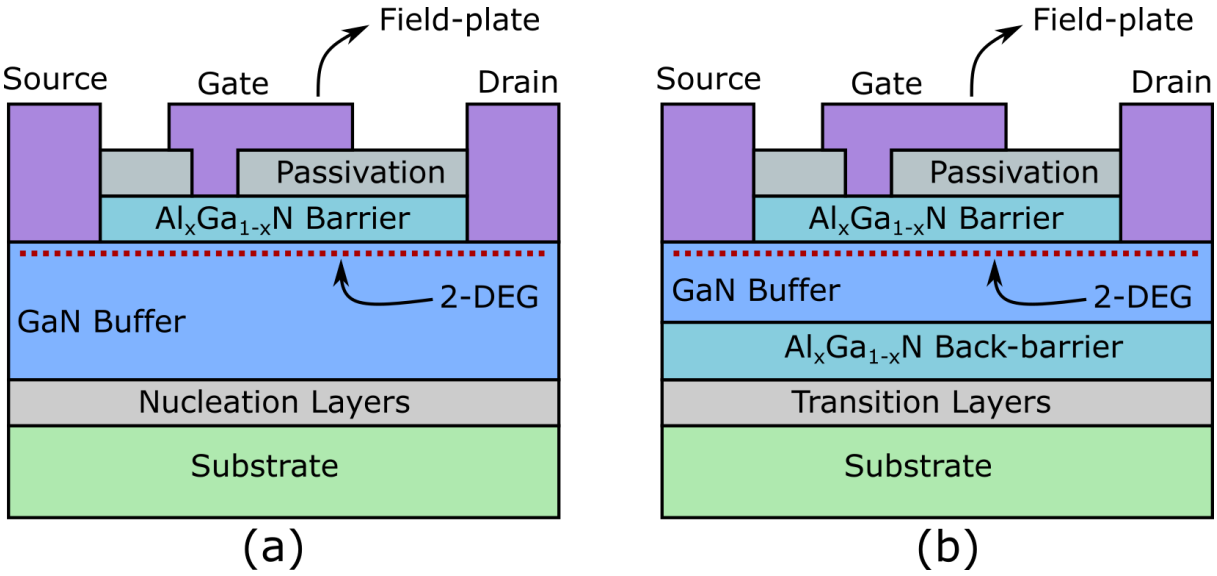


Figure 1.2: Schematic representation of GaN HEMT (not to scale). a) Traditional HEMT structure. b) Back-barrier (or DHEMT) structure.

current density in the high-field conditions encountered in PA applications. Moreover, the lower dielectric constant and the reduced geometrical dimensions for a given power reduce parasitic capacitances and delays. The largest Johnson figure of merit among all the examined materials sums up the remarkable power-frequency capabilities of Gallium Nitride [19].

1.2.2 GaN HEMT structures

Gallium-nitride devices are typically realized using a high electron mobility transistor or HEMT structure. The basic working principle of a HEMT, which found its first realization in GaAs [20], is the use of a heterojunction between different semiconductors. The junction between a wide bandgap and a narrow bandgap material produces an overall band structure that allows to create a conductive channel in an undoped region, avoiding scattering from dopant impurities and the consequent reduction in mobility.
The basic GaN HEMT structure is schematically shown in Fig. 1.2a. Due to the high-cost and technical difficulties of directly producing a high-quality GaN substrate of a sufficient thickness [21], the various layers are epitaxially grown on a different material, such as silicon carbide (SiC). The quality of this epitaxial process is of paramount importance, as defects or impurities introduced during the crystal growth can severely impact the final device performance.
The $Al_xGa_{1-x}N$ barrier layer acts as the wide gap material, as the Al fraction x can be changed in order to modulate bewtween the bandgaps observed in pure GaN ($E_g = 3.4$

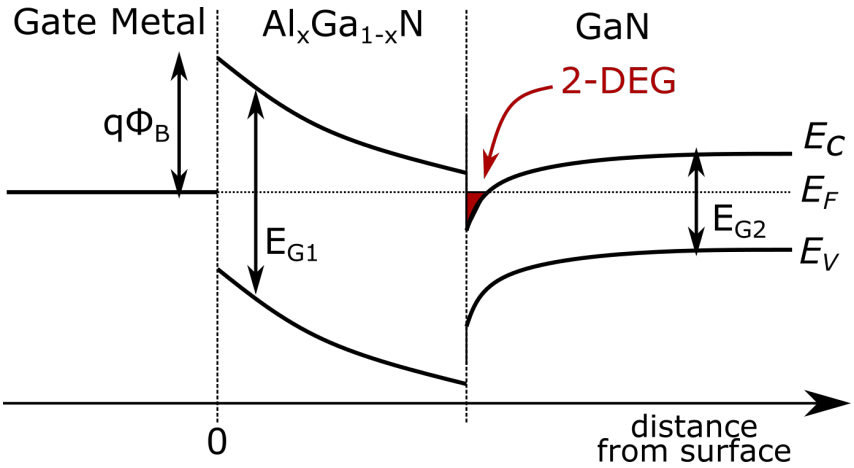


Figure 1.3: Band diagram of the gate-barrier-buffer junctions in the HEMT structure.

eV) and in pure AlN ($E_g = 6.0$eV). While other ternary compounds (e.g., InGaN [22]) or even the use of pure AlN [23] are reported, the AlGaN barrier remains by far the most used solution thanks to the relatively good matching of the lattice constant with GaN.
The narrow gap material in the heterojunction is invariably constituted by an undoped GaN buffer layer, lying below the barrier layer. The band-bending occurring at the interface between the two semiconductors causes the creation of a narrow quantum well on the GaN side, with the well energy value straddling below the Fermi level in the structure (Fig. 1.3). Therefore, free electrons can accumulate in the well, forming a high-mobility (thanks to the absence of doping in the GaN buffer) and high-density charge layer at the interface, taking the name of two-dimensional electron gas (2-DEG).
Thanks to this characteristic, HEMTs display a conductive channel even when no external voltages are applied, making them normally-on or depletion devices with a negative threshold voltage. The 2-DEG allows conduction when a voltage between the source and drain terminals is applied, achieving the required transistor effect. In order to modulate the amount of charge present in the channel region, a negative voltage has to be applied to the gate-AlGaN Schottky junction. The increase of the depletion region under this reverse bias conditions progressively reduces the amount of band bending, moving away electrons from the surface and turning off the transistor by removing the high-conductivity channel. The use of a metal-semiconductor junction in the gate contact limits the positive gate voltage swing to the small built-in potential. For larger positive values of the gate voltage, the structure displays an undesired non-zero gate current, unless an insulator is specifically introduced [24].
The formation mechanism of the 2-DEG is peculiar to GaN devices. While in AlGaAs-GaAs HEMT the 2-DEG charge comes from intentional n-type doping in the AlGaAs barrier layer, a 2-DEG with very high sheet charge density is observed in GaN devices even without any doping in the AlGaN layer. This is due to the fact that the AlGaN crystal

structure displays a strong polarization effect, and a net internal field in the material [25]. This high value of the field is enough to ionize electrons from covalent bonds, impurities or loose bonds on the AlGaN surface, which then drift towards the heterointerface and into the quantum well, forming the 2-DEG [26]. Electrons move until the the electric field is reduced, reaching an equilibrium condition. The polarization has both a spontaneous, due to the polar alignment of atomic layers when AlGaN is grown on GaN, and a strain-induced component, caused by the stress resulting from the lattice mismatch between the two materials at the interface. Therefore, thanks to this second effect, the amount of 2-DEG charge can be modified by increasing the Al fraction in the barrier layer which in turn changes the AlGaN lattice constant.
While these basic operation principles are shared by the majority of GaN HEMTS [19], a large amount of variability in the technological processes and device structures can be encountered in GaN electronics. Indeed, the optimization of the material stack [24], geometries and growth processes are still object of major research efforts in academia and industry, with the goal of maximizing RF performance. Therefore, it is not uncommon to see competing GaN HEMT solutions with radically different characteristics. As the observed effects are strongly influenced by the fine details of the underlying fabrication process, the following subsections briefly introduce some technological aspects of GaN HEMT technology.

Substrate

The choice of substrate material, on which other layers are epitaxially grown, plays a big role in the performance and the cost of the device. Indeed, the substrate has to provide high-thermally conductivity in the bulk and a good interface profile with the buffer in order to efficiently dissipate the heat generated in the HEMT active region. At the same time, the substrate lattice constant and thermal expansion coefficient should closely match the ones of the upper layers, in order to provide a low density of defects and avoid the development of stresses and cracks during high-temperature operation. The accommodation of the lattice mismatch between GaN buffer and the substrate can be eased by the use of intermediate nucleation layers (e.g., AlN), which however reduce thermal conductivity or can lead to the formation of a parasitic channel that has to be suppressed [27].
Silicon Carbide is typically the material of choice for high-performance GaN HEMT applications, thanks to the excellent thermal conductivity, the low lattice and expansion mismatch with the GaN buffer [28], [29]. However, the high final cost of this substrate, despite the availability of high-quality wafers of suitable dimensions, has pushed research to the identification of alternative, and cheaper, solutions.
Silicon is an extremely attractive substrate material thanks to its very wide availability, low cost, the large diameter wafers and the integration capabilities with existing digital platforms. However, the thermal conductivity is much lower and the lattice mismatch with

GaN is higer than SiC. Therefore, GaN-on-Si HEMT display greatly reduced power ratings and a higher amount of crystal defects during growth. Despite the technical difficulties, the development of reliable GaN-on-Si platforms is the subject of substantial research [24], [30]–[35] thanks to the prospected advantages in enabling widespread adoption of GaN technology.
Sapphire [36] has been an historically important material for the the realization of the first GaN HEMTs for high-frequency applications, thanks to its low cost and the availability of large size wafers. However, the poor thermal conductivity, large lattice mismatch and generally lower performance have led to its progressive abandonment. Instead, diamond has received renewed interest thanks to the exceptional thermal conductivity [37].

Buffer layer

The main goal of the growth of a thick GaN buffer layer is to decrease the number of defects in the 2DEG channel region, where they can negatively impact device operation. Moreover, the buffer should display semi-insulating characteristics as unwanted conduction can increase current leakage and decrease breakdown voltage. Moreover, this layer should provide good confinement of the 2DEG to the heterointerface in order to avoid injection of electrons in the buffer. Poor electron confinement has been shown to result in short channel effects that degrade transistor and PA performance, particularly in short gate length devices for sub-mm-wave applications [38]. However, intrinsic GaN buffers typically display an unintentional n-type doping due to the incorporation of impurities (su as silicon or oxygen) or the creation of vacancies and dislocations during growth [39] [40]. Therefore, different strategies have been adopted in order to implement effective buffer isolation and improve electron confinement.
The typical approach is to compensate the n-type characteristics of GaN by growing a buffer with a high density of defects or by externally adding intentional acceptor-type doping. For the latter, Carbon (C) [41] or Iron (Fe) [42] impurities are introduced at different depths in the buffer layer [43], with each solution reporting its strengths and weaknesses. However, the presence of defects and deep acceptors can adversely impact the device dynamical characteristics, as reported in Sec. 1.2.3, leading to a tradeoff between buffer isolation and dispersive effects.
An alternative solution involves the growth of an additional AlGaN layer below the channel, as shown in Fig. 1.2b, with the resulting structure being known as double-heterostructure high-electron mobility transistor (DHEMT). The AlGaN back-barrier improves electron confinement through the introduction of an additional energy barrier, avoiding the injection of carriers in the deeper layers [44].This solution would ideally remove the need of compensation doping and its unwanted effects. However, limitations in the achievable material quality and the lower thermal conductivity due to the additional AlGaN layer have limited the observed performance [40], [45], [46].

1.2.3 Trapping effects in GaN HEMTs

Despite state-of-the-art RF performance and enormous potential in power applications, GaN technology still suffers from some relevant issues. In this respect, charge trapping has been shown to be a major limitation of GaN-HEMTs, having a detrimental effect on the characteristics and reliability of these devices.
From a general perspective, a trap is an energy level located within the forbidden gap between conduction and valence bands [47]. Traps can arise from perturbations of the ideal and perfectly periodic characteristics of the semiconductor crystal lattice. The energy level is spatially localized in the region where the perturbation takes place. In a given HEMT, multiple trapping energy levels can be present either in the same [48] or different [47] positions within the structure.
There are multiple source of traps in GaN HEMTs: crystal defects due to the imperfect growth process or lattice mismatches between different layers, presence of intentional or unintentional dopant impurities, abrupt termination of the semiconductor at the surface of the device. Traps can act as capture and release centers for mobile charge carriers (i.e., usually electrons), dynamically modifying the internal charge and field distribution of the device and ultimately affecting the 2DEG density.
The level of occupation of a given trap energy level, that is, the number of trap centers that contain a charge carrier, depends on the balance between capture and emission rates. These in turn depend on a variety of quantities, such as the trap location in the bandgap, the internal temperature, the total number of defects, the number of carriers in the conduction and valence bands and the position of the quasi-Fermi levels [49].
The evolution in time of the occupation level χ of each trap can be compactly written as follows, as many of the involved quantities (e.g., quasi-Fermi level and charge densities) will ultimately depend on the applied terminal voltages V_{gs} and V_{ds}:

$$\frac{d\chi}{dt} = c_n(t) - e_n(t) = C(V_{gs}(t), V_{ds}(t), T(t), \chi(t)) - E(V_{gs}(t), V_{ds}(t), T(t), \chi(t)) \quad (1.1)$$

where c_n and e_n are the emission and capture rates, respectively. In static conditions, where there is no variation in the state of the device (i.e. all time derivatives are null), capture and emission processes balance each other, and the occupation level will be constant in time, with its value being determined by temperature and applied voltages.
Instead, in time-varying operating conditions, the variation of trap occupation levels will be generally set by the state equation (1.1). The observed dynamical evolution depends on the specific form of C and E. There are several quasi-analytic models, such as Shockley-Read-Hall theory [49]–[51], that prescribe specific functional dependencies of the terms in (1.1).
As a trend, the behavior of the two different rates in (1.1) is highly asymmetrical, with charge capture being much faster than release [47]. Capture rates strongly depend on the applied voltages (through the quasi-Fermi level [52]), increasing for higher V_{ds} and

lower V_{gs}. Instead, release rates are usually less influenced by applied voltages and show, in many cases, thermally-activated behavior, with charge de-trapping becoming faster at higher operating temperatures. The observed time constants for the trap dynamics in GaN HEMTs are typically in the μs-ns range for capture and μs to seconds or even minutes for release. In any case, the overall trap kinetics reported in experimental works is extremely complex [47], [52] and some effects cannot be described easily using existing models, particularly in heterostructure devices [53], [54].
The typical effect of traps, irrespective of the underlying physical origin, can be empirically observed as a degradation in the 2-DEG and current densities. More in detail, higher amounts of trapped charge are associated with anomalous decrease in the device current with respect to the ideal "trap-free" case. This modification of the current characteristics in HEMT in turn affects PA performance, causing a reduction of RF power [55] and unwanted memory phenomena (Sec. 1.4.1). In addition, traps have been shown to be involved in several long-term reliability issues in GaN devices [55].
The amount of trapping occurring in a device can be somewhat mitigated by refining the growth processes in order to improve crystal quality and by adopting some architectural solutions at device level. For example, since electrons in the 2DEG originate from the AlGaN barrier surface, GaN HEMT operating characteristics are strongly affected by the surface potential. Surface traps, originating from loose bonds, can capture electrons tunneling from the gate or injected from the 2DEG at large drain-source voltages, acting as virtual gates and modifying the potential in the channel [56]. The addition of passivation layers (e.g., SiN) on top of the AlGaN barrier (Fig. 1.2) greatly reduces this type of effects, while at the same time protecting the device from adverse environmental conditions [57]. At the same time, spatial extension of the gate contact called field plates [58] (Fig. 1.2) can help in the reduction of the peak electric field within the structure, which occurs at the drain side of the gate contact [40].
Despite these solutions and improvements in the technology have been successful in reducing some effects [56], significant trapping phenomena can still be observed in virtually all current state-of-the art devices. This is can be traced back to residual imperfections in the growth process and, in some cases, the presence of intentional buffer doping in order to improve HEMT performance. Given that trapping effects cannot be completely eliminated, research efforts are focused on the measurement and modeling of this type of spurious behavior.
Trapping characterization techniques have two main goals: understanding the physical origin of trapping and estimating the traps' effects on device performance. Specific techniques and models have been proposed [47], [50], [59] for each of the two tasks. At the device physical level, experimental investigations allow to identify specific defects and optimize technological processes, even though the exact link between locations and nature of defects and measured signatures is not entirely understood [47]. From a modeling perspective, instead, accurate estimations of the traps' dynamics is required in order to improve performance of the large-signal GaN-HEMT models used in RF PA design [60].

Modeling and characterization of this type of behavior are required in order to fully exploit the GaN technological potential. In this respect, the investigation of trapping phenomena in GaN HEMTs and the evaluation of its impact on RF PA design is an active research subject, with a significant part of the PhD work reported in this manuscript (Chapters 2-3) being devoted to it.

1.3 Instrumentation and measurements for RF electron devices and PAs

Empirical characterization of power amplifiers and electron devices is typically performed in a stimulus-response fashion: some excitation is applied and the consequent DUT response is observed across time. Both the stimulus and the response, in characterization for circuit design, take the form of electrical signals.
Current and voltage measurements are standard from dc up to ∼100 MHz, as suitable sensing hardware is available for both quantities in that frequency range. While active or passive voltage probes can considerably exceed the reported range, current sensing becomes impractical both for resistor-based sensors or current clamp probes. Therefore, at higher frequencies, the typical approach is to measure traveling waves instead. Their definition [61], [62] is tailored to account for distributed effects that become significant at higher frequency ranges, while still being directly related to currents and voltages:

$$\begin{aligned} a &\stackrel{def}{=} \frac{v + Z_0 i}{2\sqrt{\Re\{Z_0\}}} \\ b &\stackrel{def}{=} \frac{v - Z_0^* i}{2\sqrt{\Re\{Z_0\}}} \end{aligned} \tag{1.2}$$

where Z_0 is a normalization impedance, whose value in many typical cases is taken to be real and set to 50 Ω. The ease of measurement of traveling waves comes from the fact that the incident a and reflected b wave at a given reference plane can be efficiently separated using specific distributed passive four-port directional couplers, which take the same role that voltage and current probes have at lower frequencies. While the upper frequency limit can exceed the hundreds of GHz depending on the specific architecture, the lower frequency limit of couplers cannot extend arbitrarily close to dc due to the loss of directionality in the separation between waves.
Either in the i/v or a/b setting, the stimulus and the response signals have to be acquired concurrently in order to reconstruct the DUT behavior. Indeed, the actual excitation at the DUT reference planes cannot be taken to be equal to the theoretical one that is programmed in the digital domain within the signal source, due to the imperfect actuation in the signal generation and digital to analog conversion. If the stimulus was not measured, the resulting DUT behavior would have the unwanted consequence of embedding the non

ideal effects of the specific source used for the excitation. Therefore, the characterization of an n-port DUT has to be performed with two separate measurement receivers (one for the excitation and one for the response) for each of the n electrical ports.
The stimulus-response setting allows to take some simplifying assumptions, as the characteristics of the excitation signals will be chosen by the user and known beforehand. The excitation is typically designed to be periodic, in order to use coherent averaging of the acquisitions across multiple periods in order to lower the SNR. Moreover for many systems [63], the response of the DUT to aperiodic signals, such as the stochastic processes used to model communication signals, can be conveniently approximated using suitably designed [64] periodic excitations.
A large variety of instruments for the characterization of RF devices and PAs has been reported in literature [65], each architecture leading to different tradeoffs in terms of accuracy, capabilities and availability. One macroscopic distinction can be made on the grounds of the intended purpose. While some instruments are specifically designed for signal-level measurements, others, known as network analyzers, have the hw-sw capabilities of performing full stimulus-response characterizations of a given DUT. The development of novel characterization techniques critically leverages on a detailed knowledge of different solutions and which type of measurements are enabled by it. Therefore, due to the relevance of this implementation details for the work undertaken during the PhD, a brief description of different instruments is given in the following sections.

1.3.1 Oscilloscopes

Oscilloscopes are the standard instruments for sampling time-varying voltages, as their architecture is based around multiple fast ADCs, complemented by some preconditioning hardware in the front-end. At lower frequency ranges, they remain the most straightforward and adopted solution. The main challenge in their use for RF measurements lies in the fact that, potentially, many of the signals' spectrum components can be located at very high-frequencies (GHz-tens of GHz), and extremely high sampling rates, in order to obey the Nyquist-Shannon theorem and avoid aliasing, are required. Even if equivalent-time techniques can reduce the requirements in sampling rate [65], oscilloscopes still have the disadvantage of having to sample at double the highest frequency component in the signal instead of twice the occupied bandwidth, leading to an inefficient use of resources for typical narrowband high-frequency communication or radar signals. Moreover, the whole front-end has to handle the full frequency range of the signals and the speed requirements limit to acquisition to typically low (8-10) number of bits, reducing the overall dynamic range. Nevertheless, advances in hw technology have made extremely wideband and low-noise digital sampling oscilloscopes commercially available and used in several microwave setups [66],[67], thanks to the flexibility provided by the direct capture of the time-domain signals at all frequencies.

1.3.2 Power meters

Power meters (PM) provide measurements of incident power (i.e., rate of transfer of energy) at their input connector [68]. The input is typically well-matched to 50 Ω across their bandwidth in order to maximize power transfer to the PM, and the small residual mismatch can be compensated in post-processing. The measurements from a calibrated PM are directly traceable to SI primary power standards. The hw architectures for PMs can be broadly classified in two main classes: thermal-based or diode-based.
The first class of PMs uses a measurement of temperature variation, as the increase in temperature of the calibrated input resistor, to which the incident signal is applied, is proportional to the dissipated power. The temperature can be sensed directly using a thermocouple or as a voltage variation across a calibrated resistor bridge. As the time-constants associated to the self-heating of the resistor are much longer than the reciprocal of typical RF envelope bandwidths, thermal-based PMs can just measure the average incident power across a time-span of some seconds, without any possibility of reconstructing time-varying power envelopes.
Diode-based power PMs instead, are based on a peak-detector demodulator circuit in order to measure the envelope of the RF signal. The instantaneous incident power can then be found from the square of the amplitude envelope. These types of PMs typically display higher dynamic range but at the same time cannot work arbitrarily close to dc frequencies. As long as the RF signal is sufficiently narrowband with respect to its carrier, diode-based PMs can be used to reconstruct the power envelope across time of a dynamically varying RF signal by coupling their output with sufficiently fast ADCs [69].

1.3.3 Spectrum analyzers

Spectrum analyzers (SA) are used to measure power across frequency, that is the power spectral density, of the signal connected to their single input. Receivers in spectrum analyzer operate using the heterodyne principle: the input signal is mixed with a local-oscillator (LO) sinusoidal signal at frequency f_{LO}. For each component in the input signal at an RF frequency f, this process produces an output component at $f + f_{LO}$ and a low frequency one at $|f - f_{LO}|$. After the mixer, a narrowband filter selects just a single component of the signal, located at f_{IF}, which is called the intermediate frequency (IF). Therefore, at the output of the filter, the signal at f_{IF} will be a mapping of the components located at $\hat{f}$ in the input spectrum for which either $\hat{f} + f_{LO} = f_{IF}$ or $|\hat{f} - f_{LO}| = f_{IF}$. Several IF stages of up/downconverting mixing and filtering are typically used in order to ensure that just one frequency $\hat{f}$ in the input spectrum will be mapped at a given f_{IF}, excluding other spurious image components that can overlap on the signal of interest. The response of IF filters can typically be chosen in order to pick a different tradeoff between higher measurement speed, obtained using wide IF bandwidths, and lower noise, with narrow IF bandwidths. The mixing process, on top of shifting the frequency of signal, has

the effect of superimposing the phase of the local oscillator to the one of the incoming signal. The phase of the local oscillator changes each time its frequency is changed, due to the unknown phase locking condition of the phase-locked loops (PLL) generating the LO. While it is possible to keep track of these phase shifts in fractional-N and direct digital (DDS) synthesizers [70], this is typically not a concern in spectrum analysis applications, as just the power at a given frequency is of interest.
The power of the IF signal at the final stage of mixing is measured using with a power meter, providing an estimate of the power spectral density of the incoming signal around the single frequency $\hat{f}$. The PSD of interest is reconstructed one frequency at time on a discrete grid, where components at different frequencies $\hat{f}$ across the instrument bandwidth are measured by sweeping the internal LOs. As this sweeping process takes a finite amount of time, spectrum analyzers have limited capabilities of providing accurate acquisition of time-varying power spectra.

1.3.4 Vector signal analyzers

Vector Signal Analyzers (VSA) are used to directly characterize RF modulated signals by measuring their complex pass-band envelopes in time domain. The front end is typically composed of a single dowconverting mixing stage, possibly with IQ capabilities, followed by a wideband low-pass filter at IF and an ADC that can directly sample in an unaliased fashion the whole downconverted band. The LO frequency, even though it can be usually tuned in a considerable frequency range, is typically kept fixed for a given acquisition in order to provide a real-time demodulation of a fixed band around a known carrier, together with the corresponding time-domain evolution of the waveforms.
In this way, the envelope of given segment of spectrum can be analyzed in a phase coherent way, in a similar fashion to what a receiver would do in typical RF applications and avoiding the inefficient hw use of oscilloscopes for narrowband signals. This architecture greatly simplifies the front-end design, in which just the downconverting mixer needs to have high-frequency capabilities, while the rest of the stages can work at much lower IFs, providing lower noise and high accuracy ($\sim$ 12 bits or more). VSA software provides advanced analysis features, allowing to reconstruct the carrier of the input signals and demodulate the transmitter symbols in order to assess the signal quality metrics, such as Error Vector Magnitude (EVM), Normalized Mean Square Error (NMSE) or Adjacent Channel Power Ratio (ACPR). The real-time analysis and measurement bandwidth is ultimately limited by the bandwidth of the IF stage and the sampling frequency on the ADC, which have to comply with the standards' requirements.
In recent years, thanks to the introduction of faster, more accurate ADC front-ends and DSP capabilities, VSA functionalities have been implemented on spectrum analyzers, leading to a convergence of the architectures [65], [68]. Indeed, some of the multiple IF stages and the final power meter in a SA can be replaced using wideband ADCs, with filtering and image rejection capabilities can be implemented in the digital domain. There-

fore, similar to VSAs, modern SAs can provide some level of real-time coherent waveform acquisition capabilities.
These hybrid architectures can be augmented by combining wideband waveform capture with the LO sweeping capabilities of both instruments. For example, for periodic signals, sub-bands measured coherently with different LO frequencies can be "stitched" together in order to provide an equivalent acquisition bandwidth that greatly exceeds the ADC and IF front-end constraints. This stitching procedure [71], entails the compensation of the unknown phase shifts that occurs when the LO changes frequency. Despite the great flexibility of the approach and the possibility of providing state-of-the-art bandwidths without leveraging on sophisticated hw, its implementation on commercial instrumentation is still at an experimental stage.

1.3.5 Vector network analyzers

The standard equipment for electrical stimulus-response characterization at RF frequencies is the Vector Network Analyzer (VNA), directly providing the hw and sw means of performing system and components characterizations [68], [72]. Typical VNAs (Fig. 1.4) feature two ports, input and output in the most common implementation, but setups allowing for multiport characterizations are also available and widespread. Each port features an integrated CW excitation signal source along with a broadband RF test-set made of directional couplers and switches. The test-set separates the incident and reflected traveling waves from a DUT and redirects them to a series RF receivers. Each receiver features a single dowconversion mixing stage with a given LO, a narrow pass-band IF filter and an ADC to digitize the complex IF signal. The same LO signal is shared by all the receivers in the instrument and the final digitization stages are synchronized one to other in order to provide simultaneous acquisitions.
The basic characterization performed by a VNA entails the identification the S-parameters model of the DUT across a given range of frequencies. S-parameters provide a model of the system as a linear two-port electrical network. The linearity hypothesis allows to predict that a single frequency excitation of the device will produce a response only at the same frequency, without any spectral regrowth. Moreover, such a model is completely identified by its frequency response functions across all the bandwidth of interest.
During S-parameters measurements, the DUT is excited with a single frequency CW signal on the first port, with the second one terminated on a 50 Ω load. The scattered waves at the first and second port are coupled by directional couplers and sent to receivers. The reverse characterization is performed by exciting the second port of the DUT with a signal source, terminating the first port on a 50 Ω load and again measuring the four waves. Each term of the S parameters is identified by ratios between complex IF phasors measured at different receivers. The measured phase of each phasor will generally depend on the unknown but fixed phase shift introduced by the LO that is used for demodulation. On the other hand, ratios of coherently measured phasors at the same-frequency will not depend on the LO

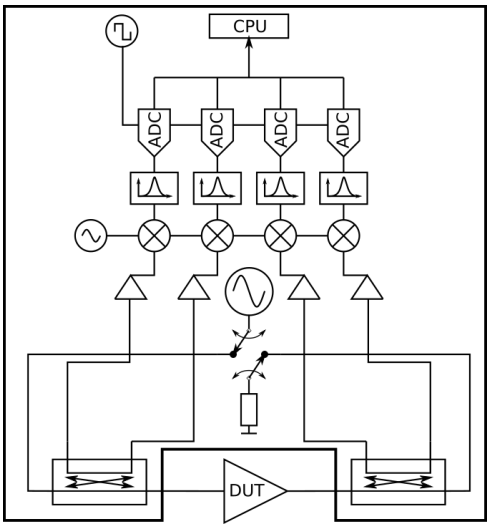


Figure 1.4: Simplified block diagram of a VNA.

phase, as it will be present with the same value both at the numerator and denominator of the ratio. The excitation frequency is then swept across the range of interest. The LO frequency is modified accordingly in order to make the incident and reflected waves at different frequencies fall at fixed IF frequency within the ADC bandwidth. In this respect, while the acquisition resembles the one in a SA, the image rejection capabilities are typically much less sophisticated, as it is known beforehand that all the measured signals will feature a single significant spectral component.
Specific vector calibration techniques are employed in order to de-embed VNA measurements from the effects introduced by the test set. As S-parameters can be expressed as ratios of waveforms, calibrations use the measurements of several known or pre-characterized passive networks in order to correct the complex-valued ratios between measurements at different receivers. After the procedure, same-frequency ratios between calibrated waves at the DUT reference planes will be traceable to fundamental standards.
Following similar trends in other RF instruments, VNAs have been adapted in order to provide characterization capabilities that greatly exceed the classical measurements of S-parameters. Using a dedicated amplitude calibration algorithm that uses a power meter as a transfer standard, a VNA can provide traceable power acquisitions. DSP algorithms allow to efficiently implement SW filtering and image rejection capabilities [73], allowing for a correct measurement of the DUT response even when modulated signals with multiple spectral lines are used as excitations. Wider IF filters and the use of faster ADCs allow to coherently capture small portions of spectrum with a fixed LO, similarly to what is done in a VSA, albeit with a much narrower instantaneous bandwidth [74]. These narrow-frequency slices can be further combined by a variety of stitching methods [71] in order to provide wider coherent characterization bandwidths and allow time-domain

reconstruction of modulated envelopes.
While the flexibility of advanced platforms can enable several waveform measurement capabilities, most of the available VNAs possess just a standard subset of these features. In this work, it is assumed that a VNA can provide vector corrected measurements of complex ratios between signal components at the same frequency. Moreover, it is assumed that each receiver can effectively operate as scalar spectrum analyzer, where absolute power levels can be measured and rejection of image and aliased components in the spectrum can be correctly performed. This still excludes any time-domain signal reconstruction capabilities, whose calibrated acquisitions entails its own set of challenges, outlined in Sec. 1.3.6. As it will be pointed out in this thesis, several advanced nonlinear characterization techniques for PAs and devices can still be performed even this reduced set of features.
Despite being usually implemented as single-box solution, VNA-like architectures can be assembled using different acquisition and test set HW, leading to several application-specific solutions [75] [66]. In this thesis, different setups will be developed using the Keysight 10 MHz-26.5 GHz PNA-X platform, which features several advanced VNA measurement capabilities and four separate test-set ports.

1.3.6 Nonlinear vector network analyzers

The measurement of time domain-waveforms in VNAs is hampered by the fact that signal components at different frequencies are measured one at a time in a non-coherent way. Thefore, in classical VNAs there will be a different LO phase shift added to each component due to the re-tuning of the PLLs between each separate narrowband acquisitions. Even if the amplitude of each component is correctly measured, the measured phases will display a random difference with the real ones, with the result of a wrong reconstruction of time-domain signals.
Therefore, to solve these fundamental issues, alternative architectures known as nonlinear VNAs (NVNA) or large signal network analyzers (LSNA) have been proposed [68], [76], [77]. Their operation can be conceptualized as reconstructing the complex amplitude and phase spectrum of the incident and reflected waves, which will be composed by a set of discrete frequency components due to the periodicity of the signals. The complex spectrum can then be Fourier transformed in order to yield time-domain representation of the waveforms of interest. Specific absolute amplitude and phase calibrations are required to correct the acquired waveforms. On the overall, NVNAs combine the wideband microwave test-set of VNAs with the measurement capabilities of oscilloscopes or advanced spectrum analyzers.
The mixer-based NVNA (Fig. 1.5a) inherits the architecture from the classical VNA, but uses an additional reference receiver. This receiver is continuously fed by a wideband periodic signal generated by the so-called harmonic phase reference (HPR). In frequency domain, the signal at the output of the HPR is composed by a series of equally spaced tones, with the tone spacing that is the reciprocal of the period of the signal. This type

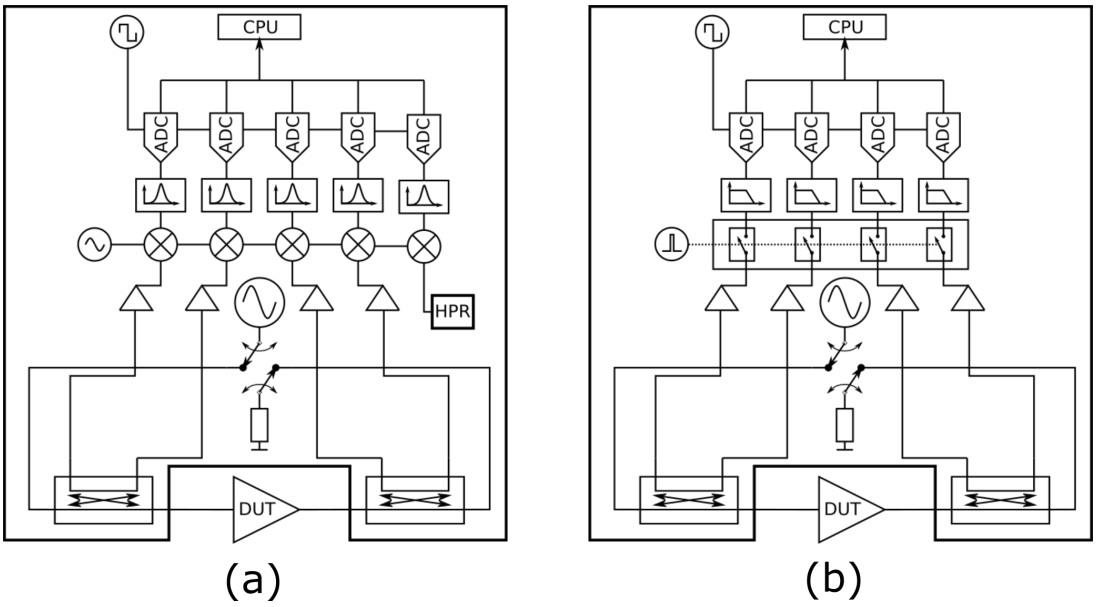


Figure 1.5: Block diagrams of nonlinear network analyzers. a) Mixer-based NVNA. b) Sampler-based LSNA.

of signal can be produced by driving a reference nonlinearity with a CW signal: the resulting distortion generates a great number of harmonics, with the amplitude and phase relations determined by the exact shape of the non-linearity. The relative phase relationships between these tones are known or are otherwise precisely characterized [78]. Signals components on the incident and reflected waves at frequencies on the HPR uniform frequency grid are measured one at a time as in a VNA. When a component is measured, the comparison between the measured phase on the reference receiver and the ideal one of the HPR allows to remove the LO phase shift from all the other receivers' measurements. Once all the harmonics are measured in a frequency sweep, they can be coherently algined to reconstruct the time-domain waveform. Mixer-based NVNAs can efficiently cover wide measurement bandwidth, but the allowed tone spacings (i.e., signal periods) are restricted, as HPRs have a limited number of significant harmonics. In order to cover RF ranges in the tens of GHz, typical commercial embodiments of the NVNA architecture feature minimum tone spacing of the order $\sim$10 MHz, which is typically too large for modulated signals of interest for the characterization of long-memory effects in PAs. Nevertheless, specific techniques can be adopted to narrow down the spacing in specific regions of spectrum [79].

The sampler-based NVNA, also known as LSNA (Fig. 1.5b), uses instead a sampling downconverter driven by a fixed frequency reference consisting of very narrow pulses in time domain. By careful choice of the pulse repetition rate, all the radio-frequency components of interest in the input are aliased, due to the undersampling process, as distinct

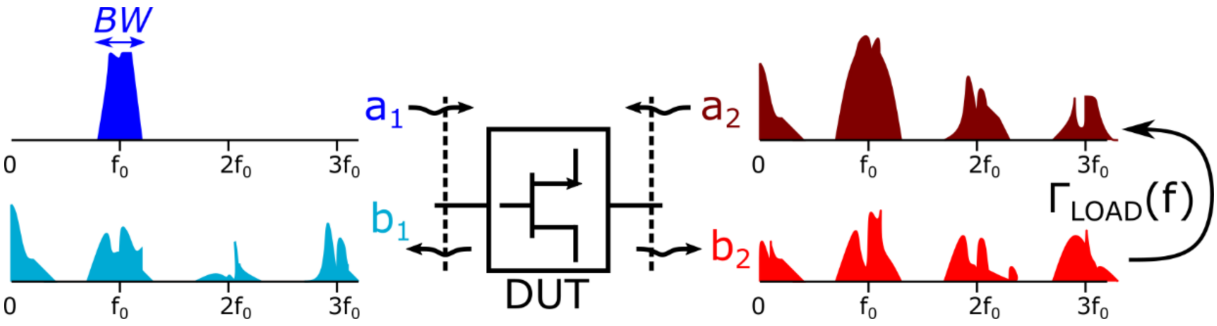


Figure 1.6: Input and output spectral regrowth components for passband modulated excitations.

low-frequency spectral lines on the output. These image components are filtered with a low pass filter and acquired simultaneously by the ADCs. DSP algorithms then "unscramble" the measured spectrum to get the real one. Here phase coherency is automatically guaranteed by the fact that they are measured in the same acquisition without requiring retuning of oscillators. LSNAs can be tailored to measure a certain bandwidth around harmonics of a given carrier, possibly with arbitrarily narrow tone spacing. While this architecture provides a good degree of flexibility, the resulting poor dynamic range with respect to other solutions led to its progressive abandonment in commercial implementations.

1.3.7 Load pull

Transistors are inherently nonlinear two-port active devices. Therefore, in general, their behavior is set by the applied signal at both ports (e.g., a_1 and a_2 waves) and they have to be characterized and modeled as MIMO networks [80].
The typical large signal characterization strategy at RF frequencies is to provide the stimulus just at a single port, while the other is terminated on some fixed passive load, constraining the overall analysis space to a single load-line linking the output variables (e.g., a_2 and b_2). This configuration replicates the circuit topology of common-source PAs which are prevalent at RF frequencies, in which the signal to be amplified is applied to the gate/base of a transistor, and an amplified version is obtained on a fixed load connected the drain/collector terminal.
The load and the excitation signal (including dc bias components) contribute to set the operating point of the DUT. Even if the excitation signal can be represented as a passband envelope around a fixed carrier in many RF applications, the spectral content of the waveforms at the device reference planes will occupy a much wider frequency support. For CW excitations, this includes dc and harmonic components, while for the modulated case, the regrowth also involves intermodulation frequencies around each harmonic, as shown in Fig. 1.6.

The analysis on a single load, however, is insufficient to characterize the overall behavior as the large signal operating point (LSOP) and the capability of active devices to provide effective power amplification strongly depends on the applied load. Indeed, a large part of PA design revolves around the choice of a specific load at the fundamental and harmonics that can optimize PA performance with respect to gain linearity, output power, efficiency and other metrics of interest [6]. It is therefore necessary to characterize the behavior of active devices as the reflection coefficient seen at the output reference plane is varied, with the technique taking the name of *load pull* [68]. VNAs or NVNAs are typically used to excite and measure the device characteristics of the devices across the load sweep.
In principle, this type of trial-and-error characterization would be unnecessary if a sufficiently accurate two-port model for the DUT was available [81]. However, in many cases, models cannot reproduce the behavior observed at different load mismatches due to the inherent complexity of the behavior of typical devices. At the same time, any complete two-port model requires identification procedures that invariable resemble some kind of load pull sweep [80]. Therefore, load pull has become a mainstay of nonlinear characterization at microwave frequencies, being used both for model extraction and for empirical optimization of device impedances.
Traditionally, load pull has been performed through the use of adjustable passive loads called *tuners*. A tuner is constituted by a slotted transmission line in which one or more screws are located. The insertion and the position of the screws can be controlled in a very precise way, either manually or using stepper motors. The magnitude of the reflection coefficient displayed by the tuner is set by the insertion of the screws, with the fully inserted case representing a short and the fully unscrewed case representing a match to the line impedance. The phase, instead, can be controlled by sliding the screws along the line. The use of multiple tuners and/or multiple screws allows to cover the harmonic case, typically up to the third order [82].
Passive load pull presents some fundamental drawbacks:

- The use of passive components inevitably entails the presence of some parasitic losses. These grow with the distance between the tuner apparatus and the reference planes of the DUT, and are typically more severe at higher frequency. This lossy behavior prevents the synthesis of impedances close to the unit circle on the Smith Chart, which might indeed be interesting in PA design [6].

- Passive tuners are inherently narrowband devices. The adjustment of the inserted screw can be used to set impedance at a single frequency point, while the load at adjacent frequencies displays a phase rotation that depends on the electrical delay introduced by the setup hw. While this type of behavior is harmless in the CW case, it is unwanted when wideband modulated excitations are used. Indeed, in the final PA, the actual load termination is typically realized much more closely to the DUT reference planes, making the introduced delay a spurious effect.

- The mechanical or manual adjustment of the screws is slow and shows poor repeatability, which is detrimental particularly in high-volume production environments. Moreover, passive tuners require a time-consuming pre-calibration step.

Despite these drawbacks, passive loadpull remains popular, particularly for high frequency and high power applications.
Active load pull, instead, synthetically emulates the effect of a passive load by injecting an additional signal on the output port. Indeed, the LSOP of the device is the same as long the the relationship between the a_2 and b_2 wave is unmodified, independently from the fact that a passive load or an externally applied excitation is emulating it. In its many incarnations, active load pull aims to solve the disadvantages of passive setups, allowing in particular to compensate for the losses introduced by the test set.
In early closed-loop active load pull setups, active injection involved the coupling of a part of the b_2 wave [83], which is then phase shifted, amplified and re-injected on the output of the DUT. By changing the amount of phase-shift and amplification, different loads can be synthesized. This configuration, however, has been shown to be quite limited in capabilities, while at the same time running into potential issues with the loop bandwidth and stability. Therefore, most of the existing active load pull solutions use an open-loop topology [75]. A second phase-locked radio-frequency source (or multiple in the harmonic case [82]) is used to inject the signal on the DUT output in order to synthesize the load, while the input excitation is kept active. The load can thus be modified by changing the injection signal characteristics, such as amplitude and/or phase. In order to synthesize a specific load, open-loop setups use a software feedback, in which the synthesized load is iteratively measured by VNA/NVNA setup and corrected until the target termination is obtained.
Wideband active load pull (WALP) techniques, extend traditional load pull capabilities by using vector signal generators as excitation sources on the input and output of the DUT. This architecture allows the maximum flexibility, as the injected signals can be fully controlled in the digital baseband domain. The main goal of WALP is to provide a target load impedance profile in a given bandwidth at fundamental and harmonics, allowing the characterization of devices and PAs under broadband modulated excitations [75]. Moreover, it allows to replicate the active termination conditions often realized in application scenarios, such as load-modulated PA architectures in Doherty [8], [84] and balanced configurations [9] or emulating antenna mismatch and cross-channel coupling in MIMO transmitters [85]. WALP capabilities can also be used to provide complete datasets for PA and device modeling [86], [87] or to significantly speed up traditional load pull characterizations [88].
Despite many favorable characteristics, active load pull setups heavily rely on the availability of high-performance microwave sources and pre-amplifiers, which can be limited in the high-power and high-frequency ranges. Moreover, the overall cost of an active load

pull system can be much higher than the corresponding passive solution, due to the use of one independent generator per targeted harmonic.
Hybrid solutions are also possible, combining both the use of a passive tuner with active injection [89] in order to compensate for the drawbacks of the two architectures.

1.4 Modeling of RF devices and PAs

Different levels of hierarchy are required in order to successfully capture the operating characteristics of a PA in a complex RF transmitter: starting from electron devices up to multi-PA assemblies. At each level, several layers of abstraction are possible in the model formulation. This choice has an impact on the model structure and on the measurement or numerical simulation tools adopted for the task.
The lowest level is the one of the the basic physical principles that regulate the system's behavior. At GaN device level, the Maxwell's and quantum mechanical electron transport equations regulate the behavior of the overall structure. These equations can be used to directly model the device electrical behavior [90] or to emulate its characteristics in T-CAD environments [91], once the material stack and geometrical characteristics are known [22]. At PA level, circuit transient or harmonic-balance simulators can represent the behavior by solving Kirchoff and components' constitutive equations [6].
However, in many cases, it is not possible to obtain a suitable model starting from physical principles, due to the inherent complexity of the system of interest. Even when possible, the complete physical simulation of a given component assembly is typically computationally unfeasible. Models are thus required in order to encapsulate the behavior of constitutive components at a given level and provide an abstraction to higher levels in the system hierarchy.
In this respect, purely behavioral or empirical modeling strategies constitute the maximum possible abstraction. These techniques have their roots in a system identification background, in which the DUT is represented as a dynamical system that mathematically relates some input and some output time-varying quantities of interest. They are also known as "black box" models, as the methods can be generally oblivious to the internal working mechanisms, but can still provide a description of the overall behavior using input and output measurements. Indeed, there is a large literature on behavioral modeling both at device [92]–[95] and PA level [96], [97].
Purely black-box approaches, however, have a region of validity that is typically very narrow, being unable to faithfully reproduce the system's behavior in different operating conditions. Therefore, a partial knowledge of the underlying physical mechanisms occurring in the device can be helpful in order to tailor measurements and tune the specific model structures. For example, while empirical equivalent-circuit HEMT models are widely used [95], their performance can be improved by adding sub-circuits for self-heating and trapping phenomena [60]. The same trend can be seen in PAs, where tailored

behavioral models targeting specific architectures have been proposed [86],[98]. This type of "grey-box" modeling tries to strike a favorable tradeoff between the generality and compactness of physical and empirical strategies.
In all cases, characterization and identification measurements play a key role. Physics-based models, despite their fundamental nature, invariably require the experimental measurement of model parameters on the actual DUT. Behavioral models, on the other hand, need specific excitation and identification strategies that can highlight the effects of interest.

1.4.1 Memory effects in RF power amplifiers and electron devices

RF devices, in order to display useful power amplification, necessarily need to exhibit some degree of nonlinearity. Therefore, the use of nonlinear models is widespread [99] at all levels of the PA hierarchy, in order to account for some of the key aspects of their observed behavior. Ideally, even a nonlinear DUT would display a static memory-less behavior, in which instantaneous output and state of the system just depend on the input at the same instant in time. While this behavior can be desirable in some cases and greatly simplifies analysis and design of RF systems [6], the absence of dynamics is typically an unrealistic assumption. The dependence of the output and state variables of a system on the past values of its inputs takes the name of *memory*.
Memory behavior is present in an RF system at several levels. For, example any PA circuit will contain capacitors and inductors to realize its functions (e.g., filtering and matching), which cause memory through charge and magnetic flux storage. These effects are also displayed by parasitic and layout components and appear, at a more fundamental level, in charge storage phenomena in semiconductor junctions in electron devices [99]. Memory is observed to affect performance is various ways, and several well-known signatures are reported in literature [100],[101].
Independently from the physical cause, which might be varied, memory effects are empirically classified in fast or slow, depending on the fact they cause the system to have short or long-range dependence from its past excitations. More in detail, for RF systems, fast-memory is characterized by time constants that are comparable with the reciprocal of the carrier frequency ($\sim$ ns and shorter), while slow effects occur on time-scales that are comparable with the modulation periods ($\sim \mu$s to s).
Fast memory typically occurs due to reactive effects introduced by circuit components (e.g., matching networks), by electron devices (e.g., intrinsic capacitances) or by parasitics in the layout. The typical signature is the variation of transistor and PA characteristics across frequency, which is more marked when the overall wideband behavior of devices and PAs is considered. Due to the great relevance of such effects in high-frequency electronics, theory and models for short-memory effects are relatively well-understood. In this, respect

they can typically be effectively modeled using variations of basic nonlinear memoryless behavior, leading to quasi-static model formulations [80], [93].
On the other hand, long memory effects are more complex to study, as their impact on RF systems is typically due to spurious behavior. Moreover, these effects can be mostly observed under modulated operating conditions, such as those in modern communication standards, because of the long-time constants at play. Several physical phenomena have been shown to cause slow memory, and both specifically tailored [102], [103] and general-purpose [101] modeling approaches have been proposed. Some of the most-commonly encountered long-memory sources in GaN RF systems are:

- *Bias networks.* Electrical components such as inductors and capacitors in the biasing circuits of electron devices display a transition from their low to their high frequency behavior. This characteristic is used to effectively decouple the signal amplification function of a PA from its connections to the power supply. The transition typically occurs in the kHz-MHz range. While the RF input envelope varies in time, the current absorbed by the PA supply terminal will display a concurrent variation. The resulting $\frac{dI}{dt}$ on the bias inductance causes an spurious dynamic modulation of the supply terminal voltage of the PA, which in turn modifies its operating conditions. In many cases, this effect can be reduced by adopting suitable circuit design techniques.

- *Self-heating.* The operation with CW signals in the RF an microwave range typically lies above the frequency at which components can effectively heat up or cool down dynamically. Instead, in modulated conditions, devices and PA then display a different amount of heating and cooling depending on the (slow) variation of the envelope waveform. As temperature strongly influences the operation of semiconductor devices, self-heating dynamically modifies the behavior of the transistors in the circuit. The effect is more severe in devices for high-power amplifiers, where the total output and dissipated power are significant in the overall energetic budget.

- *Charge trapping in electron devices.* The peak voltage applied to a GaN HEMT device varies in time when using modulated envelopes, leading to excitation of different charge trapping phenomena, such as those described in Sec. 1.2. Therefore trap occupation levels vary dynamically with the signal and influence the overall transistor behavior.

The behavioral distinction between long-term and short-term dynamics can become in some cases blurred. Indeed, the existence of effects with time-constants that vary across several orders of magnitude makes the characterization and modeling of RF systems and devices object of substantial research efforts [99], particularly in order to identify suitable measurement techniques for the task [78]. This is particularly crucial in circuits intended for modern communication standards, where wide instantaneous bandwidths (tens to thousands of MHz) are coupled with extremely long frames (in the 10 ms range or more), exciting a wide range of memory effects in an extremely complex way.

1.4.2 Volterra Series

From an empirical point of view, a nonlinear dynamical system with memory can be mathematically modeled as a time invariant operator that maps an input function of time x to another output function of time y, where both input and outputs can be vector-valued in general. Under very mild conditions satisfied by almost all practical systems of interest [104], the implicit nonlinear differential equations describing such systems can be recast, in the single-input single-output case, in the following non-recursive form:

$$y(t) = F[x(t-\tau)]_{\tau\in[0,T_m]}. \tag{1.3}$$

Therefore, the presence of memory is correctly characterized by the fact that the output at a given instant t is a generally nonlinear function of the input at the same instant and all the previous ones (and not the following ones, thanks to causality), up to an effective memory duration of T_m. If a discrete-time system is considered instead, the expression just reduces to a non-linear function of M variables, with M being the memory duration in samples.

$$y(nT) = F[x(nT), x((n-1)T), \cdots, x((n-M)T)] \tag{1.4}$$

While this general view encompasses a large variety of cases, systems of interest in characterizations for RF PA design are usually taken to belong to the periodic-in-same-period-out or PISPO subclass [63]. PISPO systems are defined by the property that a periodic input excitation will cause a periodic output response with the same period as the excitation. This corresponds to well known intermodulation behavior of amplifiers [100], in which the DUT displays multisine response to a given multisine excitation, with the resulting components falling at all the possible integer linear combinations of the original tones' frequencies. PISPO conditions can be seen to be equivalent to the physical assumptions of continuity, causality, stability and fading memory of the time-invariant operator describing the system [105]. Therefore this subclass, while still capturing a large variety of behaviors, explicitly excludes self-sustained oscillations, subharmonic generation and chaotic trajectories [63].

A large variety of behavioral models for PISPO systems have been proposed, with several parametric and non-parametric models reported in literature [99]. Among these, strategies leveraging on Volterra-theory and its modifications have been extremely popular in the RF and microwave literature [93], [106]. In this framework, for a PISPO system, the operator in Equation 1.3 can be approximated as an infinite Volterra series [105]:

$$\begin{aligned} y(t) &= y_0 + \sum_{n=1}^{+\infty} y_n(t) \\ y_n(t) &\stackrel{def}{=} \int_0^{+\infty} \cdots \int_0^{+\infty} f_n(\tau_1, \cdots, \tau_n) x(t-\tau_1) \cdots x(t-\tau_n) d\tau_1 \cdots d\tau_n \end{aligned} \tag{1.5}$$

Each term is a multidimensional convolution between the kernel f_n and n multiplied shifted versions of the input. In this respect, the Volterra series resembles a Taylor expansion of the functional in (1.3). The $n = 1$ term is the analogous of the convolution found for regular linear-time-invariant systems, for which the following equation

$$y(t) = [x * f](t) = \int_0^{T_m} f(\tau)x(t-\tau)d\tau \tag{1.6}$$

holds.
As an example, the most common LTI model for microwave devices are S-parameters. The amplifier or electron device is represented as a two-port network and the linear mapping between the incident and scattered power waves at the two ports (Sec. 1.3) is described with a 2-by-2 matrix.
While such linear models can serve as the basis for approximating nonlinear behavior either on a global (Sec 1.4.3) or local scale [107], [108], Volterra-like modeling of RF PAs and devices typically involves the identification of all the kernels up to a certain maximum order. However, the identification of higher-order terms requires the use of a large number of increasingly complex excitations [109]. Therefore, the extraction soon becomes exceedingly complex for practical applications.
Therefore, the main disadvantage of the Volterra series is that non-linear behavior is poorly described by this formulation, as the first term is just a linear convolution. The nonlinearities in RF PA and devices typically require several higher-order terms to be modeled accurately, making the direct approach unfeasible in many cases of interest. A solution to these issues can be found by a suitable equivalent reformulation of the Volterra series, called "modified" or "dynamic" Volterra series [110]

$$\begin{aligned} y(t) &= F_0[x(t)] + \sum_{n=1}^{+\infty} \tilde{y}_n(t) \\ \tilde{y}_n(t) &\stackrel{def}{=} \int_0^{+\infty} \cdots \int_0^{+\infty} F_n[x(t), \tau_1, \cdots, \tau_n] \\ &(x(t-\tau_1) - x(t)) \cdots (x(t-\tau_n) - x(t))\tau_1 \cdots d\tau_n. \end{aligned} \tag{1.7}$$

The asymptotic behavior of both the regular and modified series is equivalent, keeping the same wide-ranging descriptive power, while the order of appearance of the terms is different. The modified series makes use of nonlinear convolution integral kernels, with the first term F_0 representing a general non-linear static function of the input. Dynamic deviations from this static behavior are modeled by higher-order terms, which describe therefore the linear and non-linear memory behavior of the DUT. For many microwave systems of interest, these deviations are typically less relevant than the nonlinear static term, and a truncation of the series at first memory order has been shown to be sufficient for describing large-signal behavior in many applications [93], [106].

1.4.3 Best linear approximation framework

As discussed in the previous section, linear models have a complete theory and display several attractive properties. It is therefore reasonable to find LTI approximations for nonlinear dynamic models of systems of interest. This leads to the concept of best linear approximation (BLA). The best linear approximation of a system is defined [111], [112] as the model G_{BLA} belonging to the class of LTI operators $\mathcal{G}$ such that:

$$G_{BLA} \stackrel{def}{=} \arg\min_{G \in \mathcal{G}} \left(\mathbb{E}[|y(t) - G(x(t))|^2] \right). \tag{1.8}$$

Therefore, the BLA is the LTI system that best approximates the nonlinear system of interest, in the sense that in minimizes the mean square error between the outputs of the two systems. The BLA exists for all PISPO systems, even for ones displaying non-differentiable nonlinearities. Its value will be a combination of contributions from kernels of different orders in the Volterra series that describes the system behavior [63]. If kernels are sufficiently smooth, which is a reasonable assumption for many real systems of interest, the BLA will be smooth across frequency as well. Despite being an LTI approximation, it is in general different from the underlying linear system represented by the first Volterra kernel that can measured under small-signal excitations. At the same time, it is also different from a local linearization of the system around the LSOP of interest, which is another useful approximation used in many cases [97], [110].

In general, the value of G_{BLA} will depend on the amplitude distribution, the power spectrum and the higher order moments of the stochastic process describing the input $x(t)$. So, the BLA is the "best" approximation locally to a class of excitation signals that share these statistical properties. In typical cases [63], the input signals are taken to be stationary Gaussian processes with a user-prescribed power spectrum. The class contains as subsets gaussian noise, gaussian periodic noise and flat-amplitude random phase multisine excitations [64]. As outlined in Sec. 1.3, periodic signals are desirable in order to avoid transient non-steady state effects (i.e., frequency domain leakage) and coherent averaging techniques can be used to improve noise rejection. However, due to their limited number of degrees of freedom, they can only asymptotically approximate the statistics of the stochastic process of interest, with the approximation becoming better as the number N of tones tends to infinity. In practical applications, a number of tones in the order of $10^3 - 10^4$ [113] for random phase multitones is typically assumed to be sufficient for asymptotic BLA theory results to hold.

The use of gaussian processes, apart from the theoretical value, also covers a relevant case of interest of waveforms adopted in 5G standard specifications [3]. Indeed, OFDM waveforms [5] have a circular complex gaussian envelope for a large number of subcarriers, independently from the exact modulation format of each subcarrier. For general non-gaussian signals of interest in telecom applications (e.g., QAM modulations), the practical identification of the BLA can still be performed using suitably designed multisine excitations

[63], [113], but theoretical results are more scarse in guaranteeing statistical convergence properties.
In general, the best linear approximation will fail to predict exactly the system behavior for a specific signal belonging to the class of interest. Indeed, the BLA theory states that the measured output of any SISO system can be written in frequency domain as:

$$Y(f) = G_{BLA}(f)X(f) + Y_s(f), \tag{1.9}$$

where the first term describes the output of the BLA and the other term $Y_s(f)$ takes the name of *nonlinear stochastic distortion*. Nonlinear stochastic distortions represent the deviation of the output from the BLA due to the nonlinear distortion introduced by the DUT and displays several noise-like properties: it has zero mean ($\mathbb{E}[Y_s(f)]$), finite variance ($\mathbb{E}[|Y_s(f)|^2] = \sigma^2_{Y_s}(f)$), and it is uncorrelated with the input ($\mathbb{E}[Y_s(f)X(f)^*] = 0$) [114]. However, despite being uncorrelated, $Y_s(f)$ is not independent from the input, and its value will change depending on which particular realization of the input process is considered. Thanks to these properties, the BLA framework can be interpreted as splitting the output of a PISPO system in a part that is linearly correlated to the input and one which is not.

Measuring the BLA

The measurement of the BLA of a system of interest consists in the determination of the equivalent gain $G_{BLA}(f)$ and the variance of the nonlinear stochastic distortion $\sigma^2_{Y_s}(f)$ across the excited bandwidth of interest. To this goal, (1.9) is multiplied by $X(f)^*$ and the statistical expectation with respect to the realizations is taken:

$$\begin{aligned} Y(f)X(f)^* &= G_{BLA}X(f)X(f)^* + Y_s(f)X(f)^* \\ \mathbb{E}[Y(f)X(f)^*] &= G_{BLA}(f)\mathbb{E}[X(f)X(f)^*] + \mathbb{E}[Y_s(f)X(f)^*] \end{aligned} \tag{1.10}$$

Finally, as nonlinear stochastic distortions are uncorrelated with the input, the last term equals zero. Then, the BLA can be found as

$$G_{BLA}(f) = \frac{\mathbb{E}[Y(f)X(f)^*]}{\mathbb{E}[X(f)X(f)^*]} = \frac{S_{YX}(f)}{S_{XX}(f)} \tag{1.11}$$

where S_{YX} and S_{XX} are the input-output cross spectrum and input spectrum respectively. Conversely, multiplying (1.9) by $Y(f)^*$ and taking expectations:

$$\begin{aligned} |Y(f)|^2 &= |G_{BLA}(f)|^2|X(f)|^2 + |Y_s(f)|^2 + Y_s(f)X(f)^* + Y_s(f)^*X(f) \\ \mathbb{E}[|Y(f)|^2] &= |G_{BLA}(f)|^2\mathbb{E}[|X(f)|^2] + \mathbb{E}[|Y_s(f)|^2] \end{aligned} \tag{1.12}$$

where the input-distortion decorrelation hypothesis is used to eliminate the last two terms. Therefore, the variance of the stochastic distortions can be found as

$$\sigma^2_{Y_s}(f) = \mathbb{E}[|Y_s(f)|^2] = |Y(f)|^2 - |G_{BLA}(f)|^2|X(f)|^2 = S_{YY}(f) - |G_{BLA}(f)|^2 S_{XX}(f). \tag{1.13}$$

Both (1.11) and (1.13) need the measurements of three quantities, namely in input and output self spectral densities S_{XX} and S_{YY} and the cross spectral density S_{XY}. Several techniques have been presented in literature [63],[113],[115], leveraging different measurement capabilities and algorithms in order to provide the required measurements. Particular care has to be taken care in case of additive noise on acquisition [63] in order to get unbiased estimates of each spectrum.
In the so-called "robust" method [112] for BLA estimation, which will be used throughout this work, the measurement model in (1.9) is augmented as follows in order to account for zero-mean additive noise with variance $\sigma^2_{Y_n}(f)$ on the output:

$$Y(f) = G_{BLA}(f)X(f) + Y_s(f) + Y_n(f), \tag{1.14}$$

While input noise can cause biased estimates [63], its effect can be practically neglected as long as the input SNR of a single acquisition of the X signal is ~ 20 dB. The DUT is excited by M different realizations of a random phase multitone (Fig. 1.7), and the excitation X and response of the DUT are recorded for P periods for each realization. The resulting $M \times P$ measurements of $X^{[m,p]}(f)$ and $Y^{[m,p]}(f)$ $(m = 1 \dots M, p = 1 \dots P)$ at each of the frequencies of interest can be combined to yield the following quantities.

$$G^{[m,p]}(f) = \frac{Y^{[m,p]}(f)}{X^{[m,p]}(f)} \quad \begin{matrix} m = 1 \dots M \\ p = 1 \dots P. \end{matrix} \tag{1.15}$$

A double sample averaging of (1.15) across the periods and the realizations of $G^{[m,p]}(f)$ gives an estimation of the BLA value $G_{BLA}(f)$ at each frequency. The total variance σ^2_G of $G^{[m,p]}(f)$ across both the M realizations and P periods will be the sum of two variances, coming from independent contributions. The first accounts for noise, that varies at each period and realization, while the second is due to nonlinear stochastic distortions, which vary across realizations but are constant at each period. Instead, the variance of $G^{[m,p]}(f)$ across the P periods for a fixed realization m allows to separately estimate just the noise variance on the BLA $\sigma^2_{G_n}$ in the output measurement. By taking the difference between the total variance σ^2_G and the variance due to noise $\sigma^2_{G_n}$ [63], the variance component purely due to nonlinear stochastic distortions $\sigma^2_{G_s}$ can be found.
Finally, the variances for the Y_s and Y_n components can then be found by multiplying at each frequency $\sigma^2_{G_s}(f)$ and $\sigma^2_{G_n}(f)$ by an estimate of the input spectrum $S_{XX}(f)$.
Several different tradeoffs between the detection of nonlinear distortion and noise can be identified, by choosing a limited and fixed total number of measurements $T = M \times P$:

- $P = 1$, $M > 1$. As no periods of the same realization are available, the measured total variance will include both the measurement noise and the stochastic distortions, with no way to separate the contributions. This is not a problem if the contribution of noise is negligible with respect to the nonlinear distortion of interest, but this can cause some paradoxical behaviors at low SNRs [116].

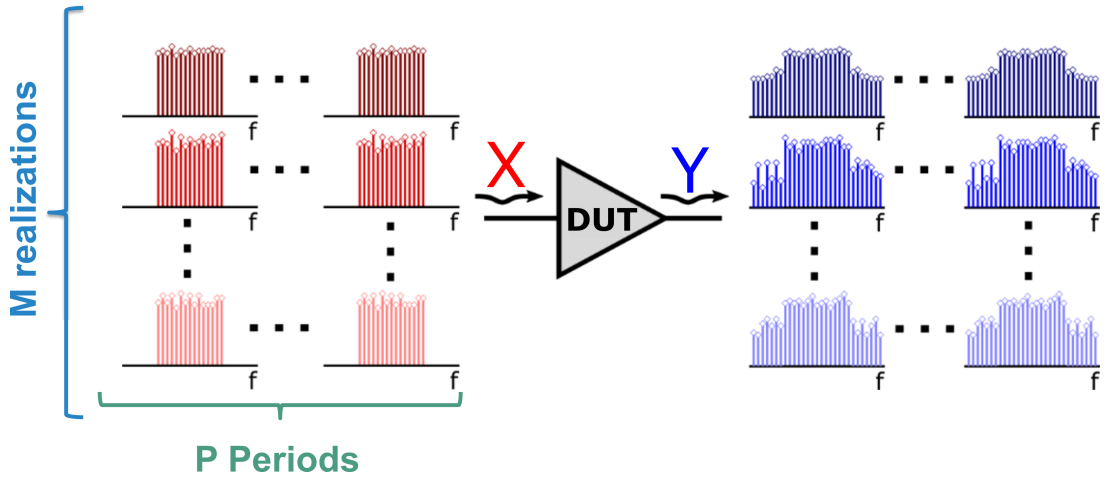


Figure 1.7: Measurement procedure for the robust estimation method of the BLA of a given PISPO DUT.

- $P = 2$, $M = \frac{T}{2}$. This configuration maximizes averaging across multiple realizations, leading to the lowest total variance of the BLA. The measurement based estimated will be closer the the "true" BLA value, as nonlinear stochastic distortion arising from independent realizations average out. Indeed, the sample means converge to the expectation operators when a larger number of measurements is taken, provided that the estimator is consistent.
- $P = \frac{T}{2}$, $M = 2$ This configuration maximizes the detection of nonlinear distortions, by decreasing the overall noise level by averaging noise across multiple periods. However, the estimate of the BLA will still contain a significant component of nonlinear distortions.
- $P = 1$, $M = 1$. At the price of lower accuracy, approximate estimation of the BLA can still be obtained using a single measurement. The technique makes use of input multitone signals that contain randomly non-excited tones in order to measure the BLA and noise/distortion variances at separate frequencies [112], similarly to what is done in noise-power-ratio (NPR) measurements [117]. Another alternative [113] involves the approximation of the expectation operator by instead averaging across nearby frequencies in the measurement, under the hypothesis that the underlying ideal BLA is smooth enough and the behavior of nonlinear stochastic distortions at different frequencies is practically uncorrelated.

Error Vector Magnitude and BLA

While the BLA has been successfully applied at device and circuit level [118], the theory has recently found renewed interest in the microwave field thanks to recent findings

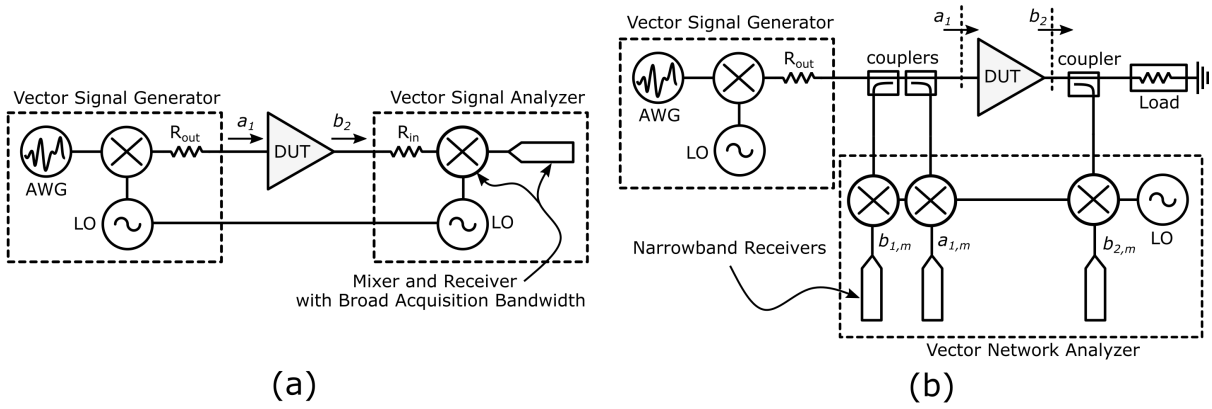


Figure 1.8: (a) EVM measurement setup based on a vector signal analyzer. (b) EVM measurement setup based on a vector network analyzer.

in its relation with Error Vector Magnitude characterization in Power Amplifiers [116]. Error Vector Magnitude (EVM) is one of the main metrics for quantifying distortion in modulated signals. EVM is defined as [119]:

$$\text{EVM } (\%) = \sqrt{\frac{\sum_{n=1}^{N} |S_n - S_n^r|^2}{\sum_{n=1}^{N} |S_n|^2}} \tag{1.16}$$

where S_n^r is the n-th received modulation symbol, S_n is the corresponding symbol in the ideal modulation constellation, and N is the number of considered symbols. From its definition, EVM is the ratio between the average error power and the average power of the ideal signal. Standard regulatory organisms [3] mandate maximum EVM values that have to be respected in RF designs. Several effects in various components of the radio-chain can affect this metric: IQ imbalance in up and downconversion mixers, distortions in the transmitter or the receiver and noise. Due to its nonlinear operation, the PA in the transmission chain is typically the largest contributor to the overall EVM. Therefore, EVM, originally a signal metric, can also be used to characterize the amount of nonlinear distortion introduced by the PA in the transmitter [116] when excited with a modulated signal.

A typical EVM characterization setup based on a Vector Signal Analyzer (VSA), which mimics the operation of radio receivers (Fig. 1.8a). The instrument filters the incoming signal within the band of interest and then automatically recovers the signal constellation by applying the required carrier synchronization and phase de-rotation. Then, thanks to the knowledge of the ideal constellation of the standard, an equalization stage allows to compensate for any linear effects introduced before the reception of the signals. In applications, this step allows the receiver to compensate for the LTI distortion (e.g, fading and multipath effects) introduced by the propagation across the wireless channel. As a

final step, as in (1.16), EVM can be mathematically calculated between the received and equalized modulated signal (S_n^r) and the reference ideal constellation (S_n). Since this procedure requires IQ demodulation and reconstruction of the time-domain signal, the VSA must feature sufficient instantaneous bandwidth (BW) with respect to the modulation in use. With the advent of 5G, the instantaneous BW requirement is up to 100 MHz for frequency range FR1 and up to 400 MHz for FR2 for a single channel, with multiple channels that can be aggregated in some applications [3]. These specifications not only involve the use of state-of-the-art VSA technology, but will necessarily pose a problematic limitation to the receiver dynamic range as the BW to be measured increases, especially for multi-channel EVM characterization over GHz-wide instantaneous BWs.
In order to overcome these limitations, various techniques have been proposed in order to characterize the EVM from the received RF waveform, without IQ demodulation [117][113]. All these techniques can be traced, in some form, to the BLA framework. Considering a DUT with input signal $x(t)$ and output signal $y(t)$, a generalized definition of EVM compares the total distortion and total signal power on the output y, assuming that $x(t)$ is the reference waveform to be transmitted. The definition, generalizing (1.16), in continuous time and frequency domain becomes (using Parseval's theorem)

$$EVM = \sqrt{\frac{\int_0^{T_m} |y_{err}(t)|^2 dt}{\int_0^{T_m} |y_{ref}(t)|^2 dt}} = \sqrt{\frac{\int_{BW} |Y_{err}(f)|^2 df}{\int_{BW} |Y_{ref}(f)|^2 df}} \tag{1.17}$$

where T_m is the measurement window and BW the bandwidth of interest for the signal. The value of Y_{ref} represents the signal portion of the output signal which is linearly correlated to the reference signal at the input $x(t)$. Linear correlation here is taken to be an LTI transformation between x and y and not a simple scalar gain g, as done in some EVM definitions [113]. In this respect, the process emulates the equalization stage in a VSA-like receiver, which estimates from the measurements the optimal equalizer filter for the signal. Instead, the Y_{err} component represents the part of EVM that cannot be linearly predicted from the input. Applying the BLA theory for the case of interest, and assuming the whole stochastic process instead of using a single realization, the power of the input-correlated component of the output is $G_{BLA}(f)S_{XX}(f)$, where the $G_{BLA}(f)$ is the best linear approximation and $S_{XX}(f)$ is the input spectral density. Then, the EVM becomes

$$EVM = \sqrt{\frac{\int_{BW} S_{YY}(f) - |G_{BLA}|^2 S_{XX}(f) df}{\int_{BW} |G_{BLA}(f)|^2 S_{XX}(f) df}}. \tag{1.18}$$

The integrand in the numerator then, using (1.13), becomes equal to the total variance of the nonlinear stochastic distortions $\sigma^2_{Y_s}(f)$:

$$EVM = \sqrt{\frac{\int_{BW} \sigma^2_{Y_s}(f)df}{\int_{BW} |G_{BLA}(f)|^2 S_{XX}(f)df}}. \tag{1.19}$$

Therefore, EVM can be estimated once the BLA characterization of the DUT is performed, for example using random-phase multisine measurements. The main advantage of this redefinition of EVM is that its measurement can be carried out without resorting to VSA and broadband demodulation HW, which, as previously outlined, entails some disadvantages particularly for wide bandwidths. Instead of a single phase-coherent acquisition, a series a narrowband ratioed measurements, such as those performed on a VNA, has been shown to be sufficient to measure correlated and distortion components [113]. Therefore, EVM characterization can be performed across the full BW of the VNA test-set, which can easily cover several GHz, and can take advantage of the noise-reduction properties of narrowband IF filters.

This approach clarifies and generalizes previous approximate EVM characterization techniques [117] that instead relied on power measurements, such as NPR, for the same goal. A typical VNA-based setup for EVM characterization is shown in Fig. 1.7b, where the incident a_1 and b_2 waves take the role of the input and output signals. Moreover, this configuration de-embeds the distortion contribution of the vector signal generator used to generate the excitation, making the measured EVM a metric of the PA only for that signal class. In all cases, this type of measurement-based EVM model for PAs requires careful design of excitations in order to approximate the stochastic process of interest [64], using high numbers of tones and/or multiple realizations to get precise asymptotic estimates of the expectations operators in the theory.

The possibility of estimating EVM without any cross-frequency phase information can also be easily seen from (1.15). The estimation of the BLA just requires the $M \times P$ measurements of the iso-frequency gain between input and output. The input power spectrum S_{XX}, needed to convert from the BLA variance to the output variance, can instead be directly measured using a power-calibrated VNA receiver. This property becomes also evident from the definition of the BLA as in equation (1.11). The power spectrum can be measured as before while the cross-spectrum expression can be recast as:

$$\begin{aligned} S_{XY}(f) = \mathbb{E}[Y(f)X(f)^*] = \mathbb{E}[|Y(f)||X(f)|e^{j(\arg[Y(f)]-\arg[X(f)])}] = \\ = \mathbb{E}[|Y(f)||X(f)|e^{j\arg\left[\frac{Y(f)}{X(f)}\right]}] \end{aligned} \tag{1.20}$$

which uses just magnitude or ratioed measurements of the quantities of interest.

1.5 This Work

1.5.1 Objectives

This Ph.D. work has focused on the development of measurement techniques for the characterization of RF GaN devices and power amplifiers. In particular, the main goal was to provide significant experimental insight into charge trapping phenomena typical of this technology, which have been shown to be both highly important in determining PA performance and, at the same time, challenging to measure and model due to their complex behavior. In order to reach a global understanding of the effects at play, the research undertaken during the Ph.D. has investigated trapping-induced dynamics at different levels in the system hierarchy: from the single HEMT, to the power amplifier circuit and the full RF transmitter.
A key aspect of all the experimental activities was the practical development of a large variety of baseband and microwave measurement benches for GaN HEMTs and PAs, in order to enable, implement and test different characterization techniques. In most cases, the work required the custom use of specific instrumentation and the design of innovative setups with advanced capabilities. As such, research in the actual measurement methods constitutes a sizable part of this work.
In this light, the scientific objectives of this PhD can be briefly summarized as follows:

- *Low-frequency current transient characterization of charge trapping phenomena in RF GaN HEMTs.* Charge trapping, due to the time-constants involved, is usually analyzed in terms of its effect on the low-frequency characteristics of the devices of interest. While standard pulsed-IV characterizations are commonly used for this task, in many cases they cannot fully describe trapping behavior, particularly given the high variability of effects between different processes. Instead, current transient measurements have the capability of shedding light on different aspects of trapping and identifying the significant time-constants at play. Therefore, one of the main research goals of this PhD was to analyze and devise specific low-frequency transient measurement strategies in order to provide a global description of trapping dynamics in GaN HEMTs.

- *RF characterization of trapping effects in GaN devices and PAs.* Despite their low-frequency nature, trapping effects strongly influence the operation of GaN HEMTs and power amplifiers under RF modulated operating conditions. During the PhD, different known and novel RF characterization techniques have been examined in terms of their capability of reproducing realistic operating conditions, highlighting the long-memory effects introduced by trapping in RF applications.

- *Development of wideband active load pull and broadband EVM measurement capabilities.* The performance and LSOP of a given DUT is critically influenced by its

load termination. This is particularly crucial in GaN devices and PAs, as trapping behavior is known to be greatly influenced by the device RF load-line and instantaneous peak voltages. In this respect, excitation and loading conditions used during experimental characterizations should closely emulate the ones encountered in the final application. For broadband modulated waveforms used in modern telecom standards, this implies the capability of controlling load impedances presented to a given DUT across a wide bandwidth. At the same time, it is necessary to evaluate the linearity performance of the DUT in the same conditions, which is a key metric in PA applications. The realization of these wideband load pull and linearity characterization capabilities generally requires the use of extremely broadband acquisitions in order to cover the spectrum of interest. In order to remove this fundamental limit, a significant part of this PhD research has focused on investigating the possibility of implementing these techniques leveraging on the narrowband acquisition hw provided by a standard VNA platform.

- *Efficient implementation of modified-Volterra models for RF PAs.* While it is known that Volterra-series based models can account for memory behavior of GaN PAs, their practical use presents several technical challenges. In particular, the nonlinear integral operators involved in the formulation make use of convolution kernels whose structure is poorly suited for simple implementation and fast simulation. Therefore, one of the PhD objectives was the study of possible efficient implementations of this type of Volterra models.

The bulk of the experimental activities has been carried out by the author in the laboratory facilities of the University of Bologna, Bologna, Italy, in collaboration with the local EDM-Lab research group. Part of the work, in particular the development of wideband active load pull capabilities, is the result of a collaboration with Keysight Technologies and the TELEMIC Group of KULeuven, Leuven, Belgium, where the author has spent 5 months as a visiting Ph.D. student.

1.5.2 Organization of the text

The present manuscript is organized as follows:

- Chapter 2 is devoted to low-frequency characterizations of charge trapping phenomena in novel short-gate-length GaN HEMT technologies for 5G applications. Different types of pulsed IV measurements are introduced, reviewing existing methods and proposing a novel technique for identifying the device's global dynamical behavior. An on-wafer measurement setup for current-transient characterizations is used in order to investigate specific trapping signatures, time-constants and energy levels in several devices employing different commercial state-of-the-art process stacks.

- Chapter 3 deals with the assessment of the impact of low-frequency trapping effects on the RF operating conditions of GaN HEMTs and PAs. Measurements are first performed at device level using a double-pulsed S-parameter technique and different large signal two-tone and pulsed excitations, analyzing the benefits and drawback of each method. Then, the characterization of a complete PA is carried out using a novel modulated multitone excitation technique, in order to closely replicate 5G-like operating conditions and analyze the system-level effects of charge trapping on broadband linearity performance metrics such as EVM.

- Chapter 4 proposes a novel wideband active load pull technique for setting a well-defined termination to a given DUT operating under realistic modulation conditions. The method allows to perform advanced load pull characterization techniques using common VNA hardware and without the instantaneous bandwidth requirements of IQ demodulation. Measurement results on a connectorized GaN HEMT are used in order to validate the approach. Moreover, the work in this chapter theoretically and experimentally investigates the possibility of extending the definition and measurement of error vector magnitude in order to perform broadband linearity characterizations in load pull conditions.

- Chapter 5 reports a novel implementation method for modified-Volterra behavioral models of RF power amplifiers. The technique uses a modified nonlinear vector-fitting algorithm, resulting in a more efficient model implementation in comparison with the existing ones. The proposed approach is validated by simulation experiments on a GaN PA.

- In Chapter 6, final conclusions are drawn. The main research results achieved during the Ph.D. are reported and some future topics and developments are proposed.

Chapter 2

Experimental low-frequency current transient characterization of GaN HEMT technologies

2.1 Introduction

The advent of the 5G standard and its FR2 implementation in the Ka/Ku-bands has recently prompted the development of advanced semiconductor technologies necessary to provide suitable radio-frequency power, linearity, and efficiency up to the millimeter-wave frequency range. In this context, given the outstanding properties in terms of RF power density, the high cut-off frequency reaching several tens of GHz, and the capability to efficiently dissipate heat, novel GaN HEMT technologies and structures are being investigated for 5G power transceivers and base stations [17].
With respect to the previous generation of GaN HEMTs, featuring 0.25-μm gate-length and mainly targeting applications in the X-band [120] and below, GaN processes for 5G FR2 applications must implement gate lengths in the order of 0.15 μm and below to provide $f_T > 35$ GHz and an effective performance in the higher frequency ranges [121].
Despite the ongoing research in process manufacturing and the attention dedicated to the 0.25-μm node during the last decade, a relatively small amount of literature data is available for device characterization, performance assessment, and behavioral modeling of sub-0.15-μm GaN devices [122], [123]. In particular, it is well-known from device physics and experience with previous technology nodes that GaN devices inherently suffer from low-frequency dispersive phenomena due to self-heating and charge trapping, and that these effects involve a complex nonlinear dependency with respect to the instantaneous voltages applied to the device ports [124], [125]. Even though commercial MMIC solutions and corresponding proprietary process design kits (PDKs) are available for sub-0.15-μm devices, many PDKs include a limited description of low-frequency dispersion due to

charge trapping and concurrent dynamic self-heating phenomena. Even for those PDKs that include such effects, specific measurement techniques are still needed in order to experimentally identify the relevant parameters that describe the trapping behavior.
The dynamical behavior, and associated time-constants, is intimately connected with the specific technological process, as different types of defects are known to display a wide range of different signatures [47], [126]. While some defects are intentionally introduced to improve the device characteristics and therefore easily predictable, others will be the results of physical imperfections in the structure that will become apparent only after manufacturing. In addition, the adoption of sub-0.15-μm gate length devices will increase the importance of short-channel effects such as poor electron confinement in presence of high electric fields [127], and the overall field-assisted trap dynamics might assume specific behaviors, depending on the device stack [128]. It is therefore of paramount importance to perform experimental characterization of these novel technologies in order to assess the most important contributions to trapping. While a completely black-box behavioral modeling strategy of these effects can be attempted [92], the best performance is observed from models that leverage on some-level of physical information obtained through specific measurements [60], [126].
The characterization of trapping effects is usually performed at low frequency [129], as the time constants in play are observed to be in the ns to s range and also evident with baseband excitations, even if the devices under examination are intended for microwave frequency applications. Moreover, trapping is typically evaluated in terms of its impact on the intrinsic IV characteristic of the device, which is of utmost importance in PA design and can be suitably measured at low frequency thanks to the negligible impact of higher frequency reactive effects in the electrical behavior.
In light of these considerations, this chapter presents the experimental low-frequency characterization results on several short gate length GaN HEMT technologies, which have been published in [54], [130]. The chapter is organized as follows: Sec. 2.2 introduces current transient techniques for the identification of charge trapping effects in GaN HEMT and describes the oscilloscope-based measurement bench developed at the University of Bologna. Section 2.3 reports the trapping characterization of a novel 100-nm GaN-On-Si DHEMT Technology. Section 2.4 proposes a novel pulsed IV method for global characterization of trapping dynamics and compares different commercial GaN HEMT technologies for mm-Wave applications in terms of trapping dynamics. Conclusions on the experimentally observed effects and their implications on the development of behavioral models and characterization techniques are drawn in Sec.2.5.

2.2 Current transient measurements for trapping characterization

2.2.1 DC IV characterization of GaN HEMTs

The measurement of the current-voltage (IV) characteristic of any electron device is the fundamental step in the assessment of their electrical behavior as a two-port network. The measurement is performed by concurrently applying dc voltages to the gate and drain port and recording the resulting gate and drain currents as the voltage values are swept in a given range. The range should lie within the safe operating area (SOA) of the device, which is delimited by a specific set of conditions in order to ensure that the device will operate reliably and not be permanently damaged during the measurement. HEMT devices are typically configured in a common-source configuration, with the source terminal tied to the potential of the bulk. Moreover, the dc gate current is typically negligible (i.e., several order of magnitude smaller) with respect to the drain one in the desired operating conditions for PA design, where the gate terminal operates as a pure voltage control contact. While even its small value can certainly provide important information on the device physics such as buffer isolation and electron confinement [131], the gate current is typically neglected in behavioral characterization in PA design, due to the reduced effect on the final specifications.
In that case, the SOA is determined by the following factors:

- Inverse polarization of the gate Schottky junction: if the metal-semiconductor interface becomes directly polarized, the gate terminal will start to conduct a non zero current, denying the voltage control conditions on the gate. The avoidance of this condition imposes that the gate applied voltage should typically be less than 0 V or a slightly positive value, up to the diode threshold voltage.

- Breakdown: the effect is marked by a sharp increase in the drain current for large drain voltages and can be attributed due to a variety of physical effects [132]. While the condition is in generally reversible and not necessarily destructive with the due precautions, the device will no longer operate in the ideal voltage-controlled conditions typically required for amplifier applications. This imposes a maximum drain voltage that has to be applied during IV characterizations.

- Self-heating: At higher temperatures, HEMTs are known to experience a variety of performance such as increased gate conduction, premature degradation phenomena, reduced breakdown voltages and lower reliability [36]. Therefore, the instantaneous power dissipation given by the product between drain voltage and current and subsequent self heating must be kept under control, depending on the thermal resistance of the device.

The main disadvantage of this dc IV characterization of devices is that the measured characteristics will not be isothermal and therefore will embed significant self heating. Indeed, each point on the IV curves will generally feature a different dissipated power, a different amount of self heating and a different temperature distribution in the device active area, even if the baseplate temperature is the same for the whole characterization. As known from basic device physics, semiconductor devices display a strongly temperature-dependent behavior [36] and such dependencies have to be carefully distinguished from the actual electrical characteristics that are needed for circuit design.

2.2.2 Pulsed IV characterization of GaN HEMTs

In order to avoid these shortcomings, pulsed IV (PIV) measurements have been introduced [133]. PIV employ two-level periodic square-wave excitations on the gate and drain of the DUT. The two levels are respectively called the pulsed voltage, which is applied for a sufficiently short fraction of the period, and the quiescent voltage, applied for the largest part of the period. The quiescent voltages are typically chosen in order to ensure zero current and therefore zero dissipation conditions, so that the internal temperature will equal the one imposed by the external baseplate. Another common choice [69] is to pick a quiescent point that closely matches the one intended for the operating class (e.g. AB, C or others) for the PA in which the device will be used, in order to partially reproduce the final application conditions. Instead, the pulsed voltages are swept in the SOA as in regular IV characterizations.
If the pulsed voltage is applied for a time shorter than typical self-heating time-constants (in the μs-ms range), the device temperature will not substantially depart from the value at the quiescent conditions. Any slight temperature deviation is recovered during the much longer quiescent period. The current observed during the pulse will be practically constant in time once electrical transients due to reactive effects in the pulsing instrumentation have died out [134]. Its average value can be taken as representative of the current of the DUT under the pulsed voltages. Therefore, the average pulsed current can be measured as an independent function of the pulsed voltages and the baseplate temperature, yielding isothermal IV characteristics.
Another advantage of PIV characterizations is that they can greatly extend the dc SOA, as the device operates in a practically isothermal fashion without any additional self heating. Indeed, the device can temporarily withstand conditions outside the dc thermal SOA for a short amount of time, without incurring in irreversible degradation or reduced reliability that would be observed if the same fixed voltage values were applied for a long (at the limit infinite) amount of time. This fact is exploited by PA designers, who employ dynamic loadlines that can briefly exit the dc SOA in order to maximize current and voltage swings for increased device performance. Thanks to their capability of fully covering the operating region of the PA, isothermal PIV form the basis of most equivalent-circuit model formulations used in designs.

Ideally, given these premises, the measured PIV characteristics of the device should be the same even when the quiescent state is moved across the zero current region of the device, provided that the correct timing relationships are maintained. However, this is not the case if trapping effects are at play, such as in GaN HEMT devices. As discussed in Sec.1.2, the trap occupation levels and transition rates vary with applied voltages, with higher drain and lower gate voltages related to higher amounts of trapped charge [49]. Moreover, traps display highly asymmetrical dynamics, with capture rates being several order of magnitude slower than emission ones and showing different temperature and voltage dependencies [52], [135]. Therefore, even the short pulses applied in PIV are enough to increase the amount of captured charge, while being too short for any significant release to happen.
The deviation between PIV characteristics measured from different quiescent points can be used to assess the amount of trapping that is taking place in the specific DUT [60], [107], as different amount of trapped charge during the pulsed conditions will generally produce different currents. Some relevant test cases using different off-state quiescent points have been reported in literature [60], [136]:

- No applied quiescent voltages. In this case the trap occupation state during the quiescent state will be at the equilibrium level. Therefore, any pulsed voltage will generally modify the trapping state with respect to the quiescent condition and the observed current will be indicative of the trapping state during the pulse itself.

- Gate Lag. The applied quiescent voltage is taken to be zero on the drain and minimum (i.e., highly negative) on the gate. This quiescent point enhances trapping observed due the highly negative gate voltage, without affecting the drain, which has been shown to mostly affect traps located at the surface of the barrier layer [56].

- Drain Lag. The applied quiescent voltage is taken to be maximum on the drain and minimum (i.e., highly negative) on the gate. The amount of trapped charge during the quiescent conditions is at its maximum (for the given SOA): on top of the traps charged during the gate-lag characterization, the high drain voltage has been shown also to promote trapping in the GaN buffer region of the device [136]. As the pulse width is typically much shorter than the charge emission time constants, the current measured during the pulse will be indicative of the trapping conditions in the quiescent state, as no significant detrapping can take place during the pulse. Therefore, the resulting PIV characteristic is measured with a fixed maximum trapping state.

Intermediate situations can be observed by picking other points in the device SOA. What is generally observed is that PIV characteristics from "high-trapping" points display, for a given gate and drain voltages, a reduced current level with respect to the PIV measured from the zero voltage point. This phenomenon is known as *current collapse* [137], which is usually accompanied by the related *knee walkout*, that is, an increase in

the knee voltage of the device. While trapping effects are efficiently identified using by gate and drain-lag PIV measurements, their presence can be detected even from in standard dc IV characterizations. In that case, trapping is phenomenologically as the measured current displays anomalous dependencies on the applied voltages, such as "kinks" [138] or increased gate leakage in the reverse region [23], with respect to the characteristics predicted from device physics and TCAD simulations.
In order to improve on the basic PIV characterizations in providing isothermal and iso-trapping device characteristics, Double Pulsed IVs (DPIVs) have been introduced [107]. In DPIV, the quiescent point is chosen in the device active region that is typically used for PA design, such as a class AB bias point. The usual pulse is preceded by a similar duration pre-pulse towards the drain lag point. Given that trapping time constants are expected to be shorter than typical pulse durations, the pre-pulse preconditions the trap level to their maximum occupation state. The pre-pulse is immediately followed by another measurement pulse towards some point in the SOA, like in classical PIV. In this way, the current measurement is performed in a well-defined and fixed trapping and thermal state [139], as the measurement pulse is designed to be much shorter than detrapping and thermal time constants. While also drain lag characterization is expected to provide iso-trap IVs, the DPIV has been shown to more closely reproduce the current-voltage behavior that is displayed when using dynamic excitations in PAs [140]. In that case, depending on the actual waveform, the dynamic loadline is mostly contained within the active region, with crossings in high-trapping (i.e., high-drain, low-gate) areas for brief amounts of time. Instead, in drain lag, almost all the time is spent in the high-trapping region, yielding an overestimation, for design purposes, of the current collapse in the device.

2.2.3 Current transient characterization, I-DLTS and Arrhenius analysis

PIV or DPIV characterizations are a fundamental step in identifying the effects of trapping behavior on the electrical characteristics of the DUT. However, the obtained information is partial, as any PIV characteristic is just a static snapshot of a much complex capture/release dynamical evolution (Sec. 1.2), without any clue on the precise time-scales that are at play. Moreover, the choice of pulse parameters, such as duration of the pulse and pulse periods, is critical in determining the physical significance of the observed results. This is due to the fact that all these techniques rely on some physical assumptions (e.g., fast trapping, slow release) that have to be precisely quantified. In this respect, values that are suitable for a given technology, might not be valid for another due to the large variety of effects [102].
For example in DPIV, even if capture is thought as an extremely fast process, in some cases this assumption has been shown to be violated, with a significant increase in the

trapped charge being observed for longer pre-pulse durations [141]. Detrapping transients instead can show time-constants that vary across several orders of magnitude depending on applied voltages and temperature [47] and typical periods for PIV can be too short to observe the complete extinction of such transients. In the general case, there will be a complex interaction between the pulse timings and the trapping time constants, which can impair any effective behavioral comprehension of the device and separation of various effects at play.

Therefore, the detailed knowledge of the dynamics of the charge capture and release effects is necessary to perform characterization of GaN HEMTs for PA design. As discussed, even the measurement of basic PIVs requires some level of knowledge of these effects in order to obtain any meaningful insight. Moreover, results are expected to be highly DUT dependent and a "one-fits-all" approach is hardly applicable. This is a major challenge for behavioral modeling and design efforts, which instead greatly benefit from technology agnostic approaches.

To this goal of gaining an a-priori understanding of the charge capture and release dynamics in a given device, a large variety of methods have been proposed in literature [47]. These techniques have originally been devised in the field of semiconductor materials characterization, where the main interest is identifying key physical properties of defects in the HEMT structure. While physics-based methods such as photoluminescence spectroscopy [142] can be used, the experimental characterization of such low-frequency dynamic effects in GaN HEMTs is usually achieved by using different types of electrical measurements [123].

One possibility involves small-signal (Y-parameters) [91] or noise [123] measurements in the dispersion frequency range. Trapping/detrapping time constants can easily be identified as peaks in the resulting frequency response functions, but typical biasing constraints often restrict the characterization to a limited part of the SOA. Moreover, from a behavioral perspective, the adopted small-signal regime does not easily allow to characterize the global nonlinear dynamic dependencies on the applied voltages.

An alternative method consists in the analysis of capacitance [143] or current [144] transients after the application of specific voltage steps, in a technique called deep-level transient spectroscopy (DLTS). In particular, the current-mode DTLS (I-DLTS) has been extensively used for trapping characterizations in GaN HEMTs [47]. The basic incarnation of the technique involves the application of a filling pulse that increases the occupation level of traps. The pulse can be applied either on the drain, the gate or both, eventually with the use of specific load resistors to emulate application-specific loadlines [145]. After the pulse, a given quiescent voltage is applied, and the resulting current transient to lower trapping condition is acquired until equilibrium is reached. The variations in transient responses are observed by varying pulsed and quiescent voltages, temperatures and pulse widths, with specific behaviors that can be used to reconstruct the physical cause of the trap, such as point or extended defects, dislocations or intentional doping in some of the HEMT layers [47], [52].

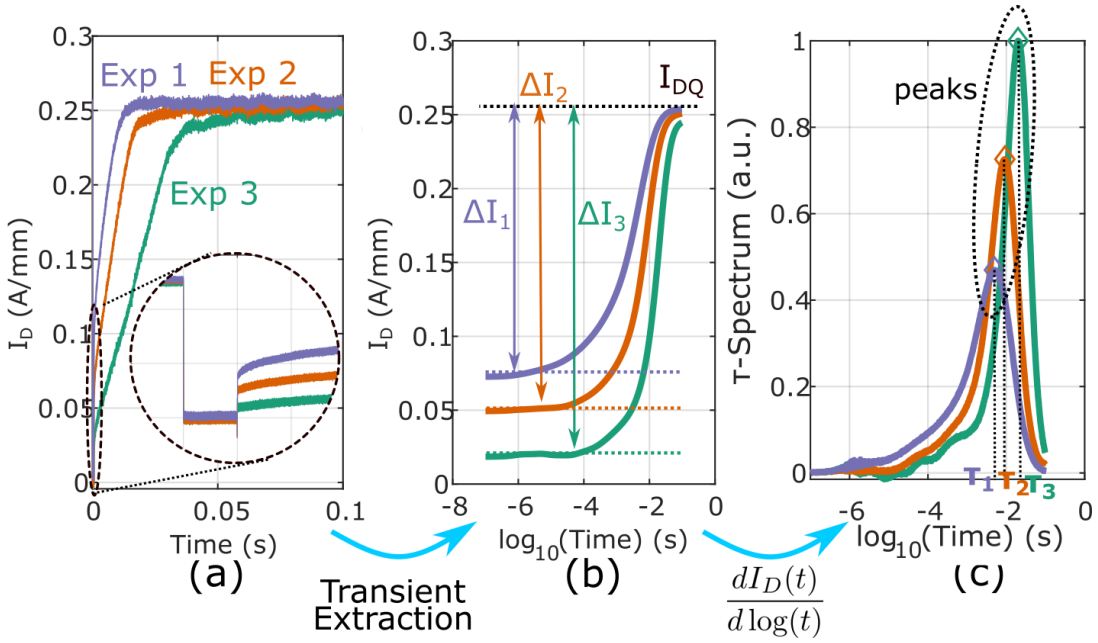


Figure 2.1: Procedure used for the drain current transient measurements. (a) Current transient acquisition. (b) Plot of the current transients in logarithmic time scale and evaluation of the current drop. (c) Log-time derivative and extraction of trap time constants and peak values.

In I-DLTS the instantaneous current value and its variations are used in order to probe the time-evolution of the inaccessible internal trap occupation levels. As it will be discussed in Sec. 2.2.4, measurement setups for this technique closely resemble the ones used in PIV characterizations, but should provide wideband waveform capture capabilities and large samples memories in order to characterize dynamics across a large span of time-scales. With respect to other techniques, the use of the drain current as a variable of interest provides direct behavioral information on the impact of trapping on the IV characteristics of the device, which is crucial in PA design application. On the other hand, no information can be obtained in zero-current conditions, where the trapping state has to be indirectly inferred using other methods.
In general, current transients (Fig. 2.1a) after a trap-filling pulsed stimulus feature complex de-trapping dynamics, described by a continuous spectrum of stretched exponential relaxations, which can be estimated using a wide variety of methods [28], [47], [143], [146], [147]. During the research performed during the PhD, the popular method in [147], which is based on the log-time derivative of the current transient $\frac{dI_D(t)}{d\log(t)}$, is adopted as a representation of the time-constant spectrum for a given excitation (Fig. 2.1c). In this description, the presence of a strong peak (i.e., local maximum or minimum) is a key signature of the presence of a relevant trapping energy level, whose characteristic time-constant can then

be deduced as the time location of the peak. Multiple peaks or even negative peaks (i.e. negative derivative and decreasing current transients) related to hole traps have been observed for some GaN HEMT technologies [47], [48], indicating the existence of multiple trapping mechanisms in the structure, each with a different time-constant.
The identification of these characteristic time constants also allows to identify the corresponding trapping energy levels, which can shed light on the physical origin of the observed trapping phenomenon such as distinguishing between defects at interfaces, surface states, and intentional or unintentional doping in the buffer layers [47], [147]. This is done by studying the variation of the emission time constant with respect to temperature. Following Shockley-Read-Hall theory, for a single electron trap with an energy level located E_a below the conduction band edge E_c, the electron emission rate e_n has the following relationship with temperature T, which is called Arrhenius equation [148]:

$$e_n = \frac{1}{\tau_n} = \sigma_n A T^2 e^{\frac{-E_a}{KT}} \tag{2.1}$$

where σ_n is the capture cross section of the trap, A is a temperature-independent factor containing several semiconductor material parameters [22] and k is the Boltzmann constant. The Arrhenius equation predicts that charge release processes typically depend strongly on the temperature, with the time constant getting progressively shorter for increasing temperatures. Therefore, the practical procedure to identify the activation energy E_a involves the measurement of I-DLTS transients by varying the baseplate temperature. The effective temperature need also further compensation in order to account for device self-heating [149]. Finally, (2.1) can be recast in the following form:

$$\ln(\tau_n T^2) = -\ln(\sigma_n A) + \frac{E_a}{kT}. \tag{2.2}$$

The activation energy can then be found as the slope of the best-fit line when the temperature/trapping-constant pairs are plotted with coordinate axes equal to $\ln(\tau_n T^2)$ and $\frac{1}{kT}$ respectively, with the resulting graph known as Arrhenius plot.
While the SRH model succeeds in explaining the detrapping kinetics in many cases [150], [151] and correctly accounts for self-heating effects that are sometimes neglected in reduced equivalent-circuit models [152], it can be oversimplified in many other relevant situations. Indeed, trapping can be associated to extended instead of point defects [153] and charge release can occur through non-purely-thermal mechanisms, such as field-assisted emission [146] or hopping and tunneling [47]. Therefore, an Arrhenius characterization and activation energy extraction is an important, yet non-exhaustive, step in investigating the relevant trapping processes in a given technology.
Other relevant information that can be obtained from I-DTLS characterization is related to the amplitude of the current transient [154], which can be traced back to the severity of the current drop in PIV characteristics. Typical ways to quantify are normalized current differences between the beginning and the end of the transients (Fig. 2.1b) or the peak

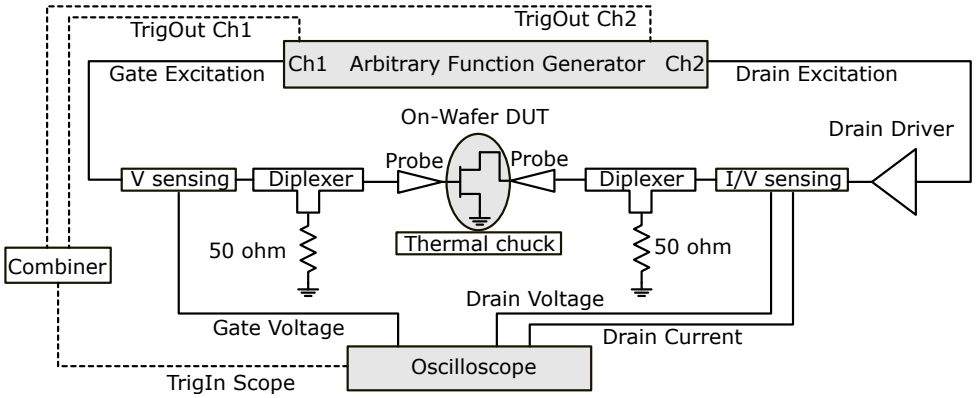


Figure 2.2: Block diagram of the measurement setup (10-MHz reference not shown).

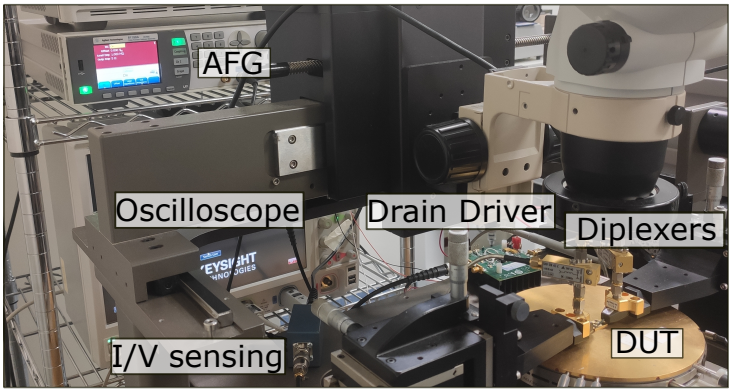


Figure 2.3: Photo of the on-wafer measurement setup.

current derivative in I-DTLS traces (Fig. 2.1c). I-DLTS has typically been used in order to characterize charge release transients, while capture transients have been rarely studied in literature [155]. The two main causes are that these transients are typically much faster than the capabilities of many measurement setups and that typical pulsed voltage configurations that induce trapping also induce an almost-null pulsed drain current, which prevents the study of the trapping evolution. Nevertheless, Sec.2.3 will also investigate the possibility of performing this type characterizations.

2.2.4 Measurement setup

In order to enable the trapping dynamics characterization capabilities introduced in Sec. 2.2.3, the low-frequency measurement setup first proposed in [107] has been fundamentally revised and updated. The developed on-wafer bench is shown in Figs. 2.2-2.3. With respect to the previous incarnation, the new setup allows greater flexibility in the design of experiments, eliminating the requirement of maximum waveform repetition period of

100 μs and allowing the use of arbitrarily long pulse periods and widths. This extension down to dc provides the means for exciting, acquiring and analyzing long transient effects. The corresponding reduction in available voltages, due to limitations in the new HW, is less of a concern for short-length devices for mm-wave applications that were examined during the PhD.
The devices-under-test are on-wafer GaN HEMTs in common source configuration, where gate and drain pads are accessed by 150-μm-pitch ground-signal-ground (GSG) probes. A dedicated thermal chuck allows the set and control the baseplate temperature of the GaN dies inserted in the probe station.
The excitation consists of different periodic pulsed voltages concurrently applied at gate and drain ports with a two-channel, 120-MHz arbitrary waveform generator by Agilent (81150A). The gate port can be directly controlled by one of the AWG channels without any additional conditioning, thanks to the reduced voltage and current levels required. At the drain port, an analog driver based on a wideband current feedback amplifier (Analog Devices ADA4870), featuring up to 50-MHz LS BW with I/V swings of up to 1 A and $\sim$38 V, is adopted for handling the electrical characteristics of the DUTs. The pulsed excitations are applied through the low-frequency path of dedicated connectorized diplexers (SHF DX45) with 25-MHz nominal BW (dc coupled) on the low-frequency port, while the RF ports of the diplexers are terminated to 50 Ω to enhance device stability. Differently from other setups [134], this configurations avoids the use of narrowband bias-tees, which may influence the device terminations and introduce ringings, whose instantaneous voltage peaks can jeopardize the characterization. At the same time, it keeps the possibility of using arbitrarily long excitation periods to characterize the dynamics of slow detrapping processes, which is unavailable in other setups which use the ac side of bias tees to applied the pulsed voltage excitations [156]. The resulting wavefronts have rise and fall time in the order of 100ns, limited by the cutoff of the diplexers and the LS BW of the drain driver.
The dynamic voltage waveforms are measured with commercial passive 1:10 1 MΩ voltage probes, and the drain current is acquired with a 100-MHz current clamp probe (Keysight N2893A). All the probes are connected to a 2.5-GHz BW, 20-GSa/s digital sampling oscilloscope (Keysight 9254A). This configurations allows for wideband pulsed-waveform acquisition as well as sufficient data memory to capture ms-range pulse periods at a sampling rate of 100 MSa/s, covering up to 50 MHz of acquired unaliased analog bandwidth with a resulting samping time of 10 ns. Hence, the setup allows to finely capture not only the slow de-trapping transients, but also the fast capture transients which often fall beyond the BW available in typical pulsed setups. Even if the present chapter will focus on pulsed waveforms with different characteristics, the setup has the flexibility of applying any arbitrary concurrent voltage excitation on the two ports, enabling a large variety of low-frequency characterizations.

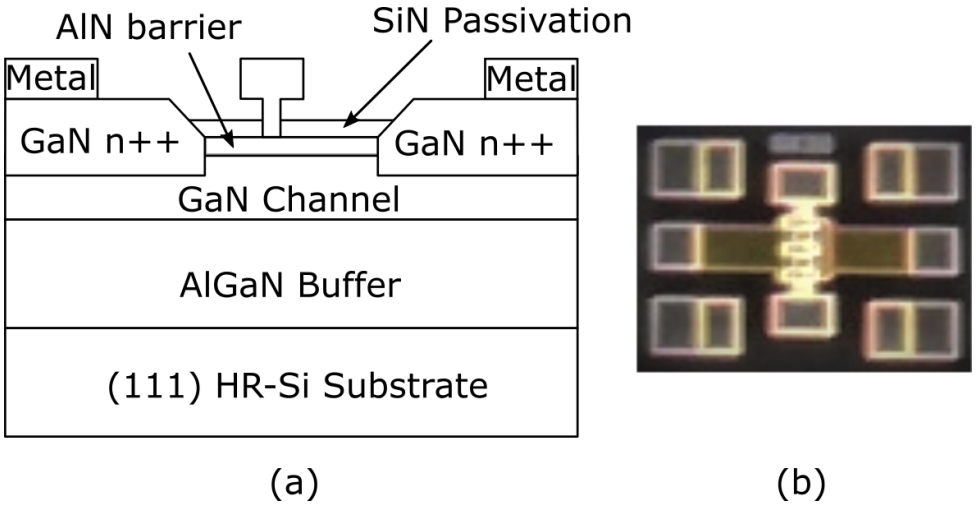


Figure 2.4: GaN-on-Si technology: (a) HEMT structure. (b) Die photo.

2.3 Charge trapping dynamics in 100-nm AlN/ GaN/ AlGaN-on-Si DHEMT technology

2.3.1 Process technology

The first examined DUT is a 100-nm mushroom-gate double-heterojunction (DHEMT) AlN/GaN/AlGaN HEMT with 250-nm gate-source distance. This novel GaN-on-Si process is designed to overcome the typical AlGaN/GaN limitations for sub-200-nm gate lengths [157], where the punch-through [38] of the buffer layer in the presence of high electric fields may induce short channel effects. As shown in Fig. 2.4a, the epitaxial structure is grown on (111) high-resistive silicon (HR-Si) substrate. The introduction of an AlGaN layer acts as a back barrier preventing the electron flow into the buffer under high drain-source voltage (50 V breakdown), improving the electron confinement and allowing for high drain current density (1.3 A/mm at $V_{DS} = 3$ V) without the intentional introduction of buffer dopants [23], [131]. Due to the high spontaneous polarization and the wide-bandgap (6.2 eV) of the AlN barrier, the AlN/GaN heterostructure effectively allows for high 2DEG density ($> 10^{13}$ cm^{-2}) [158]. In-situ grown SiN layer minimizes the strain relaxation, reducing the defects on the AlN layer and improving the reliability of the device under high electric fields [57]. The resulting cutoff frequency is f_t=120 GHz. At 30 GHz and $V_{DS} = 12$ V, the typical RF power density is 3.3 W/mm (6.6 W/mm peak), with a maximum stable gain of 13 dB (2x25 μm device). The device-under-test (DUT) is a 6×30 μm HEMT die in common-source configuration (Fig. 2.4b). The threshold voltage measured at dc conditions is $V_{TH} \approx -1.6$ V for a current density of $I_D \approx 5\frac{\text{mA}}{\text{mm}}$. This value displays almost no dependency on the applied V_{DS}, thanks to the AlGaN back-barrier

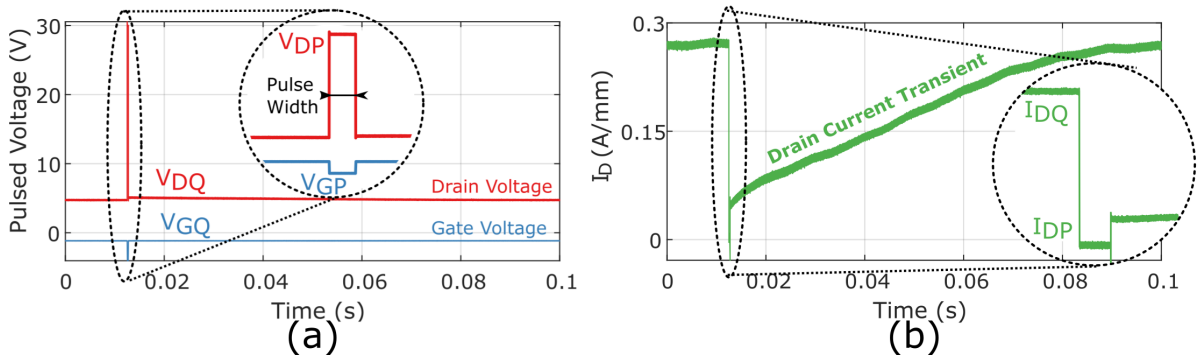


Figure 2.5: Actual waveforms acquired with sampling time of 10 ns for concurrent gate and drain pulsed excitations inducing charge trapping and causing a drain current transient recovery.

preventing significant punch-through effects [38].

2.3.2 Configuration of the pulsed excitations

The excitation consists of voltage pulses concurrently applied at gate and drain ports. The periodic pulsed waveforms are defined by a period $T = 10^{-2}$ s (one period per acquisition), and pulse widths $PW \in 5 \times [10^{-8}, 10^{-4}]$ s. The pulsed voltages applied during PW, i.e. labeled as V_{GP} (gate) and V_{DP} (drain), while the voltages applied during the $T - PW$ time-window (baseline) are referred to as quiescent voltages V_{GQ} (gate) and V_{DQ} (drain). An example of actual acquisitions is reported in Fig. 2.5, showing the capabilities of the setup to deliver clean pulsed voltage waveforms.
In Fig. 2.6, we report the PIV characteristics from the nominal quiescent point, in comparison with the ones pulsed from $V_{GQ} = 0$ V, $V_{DQ} = 0$ V and $V_{GQ} = -2$ V, $V_{DQ} = 20$ V, exhibiting the presence of significant lag effects, and a trap-induced current reduction in the order of 100 mA/mm coupled with a significant knee walkout.
In general, a certain trapping state will be set depending on V_{GP}, V_{DP}, PW and T, which can be studied by analyzing the long current transients recovering to quiescent conditions (Fig. 2.5) after different pulsed configurations. To this aim, in Figs. 2.7 we report a representative characterization, in which the pulse amplitudes cover the four possible relationships between pulsed and quiescent values. In this way we can explore different situations, in which either capture or release occur during the pulsed or quiescent conditions.
The results n Fig. 2.8 confirm that the most evident current drop and longer recovery, hence the larger amount of trapped charge during PW, takes place when $V_{GP} < V_{GQ}$ and $V_{DP} > V_{DQ}$ (Fig. 2.8d). This is also the most evident behavior for LS PA design, as the dynamic RF loadline will typically reach maximum trapping conditions [139]. Conversely,

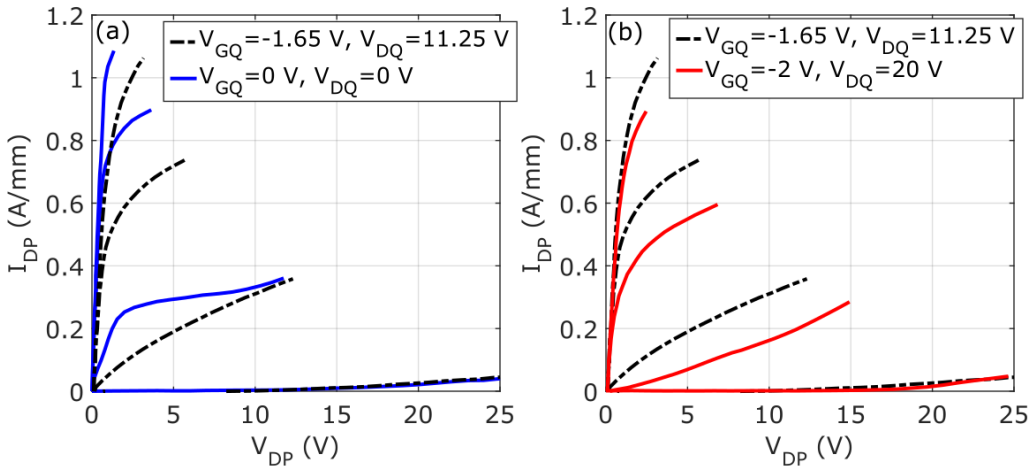


Figure 2.6: Pulsed-IVs from quiescent points (V_{GQ},V_{DQ})=(0 V, 0 V), (-2 V, 20 V) and (-1.65 V, 11.25 V). V_{GP} from -1.8 V to 0 V with 0.6 V step. PW=100 ns.

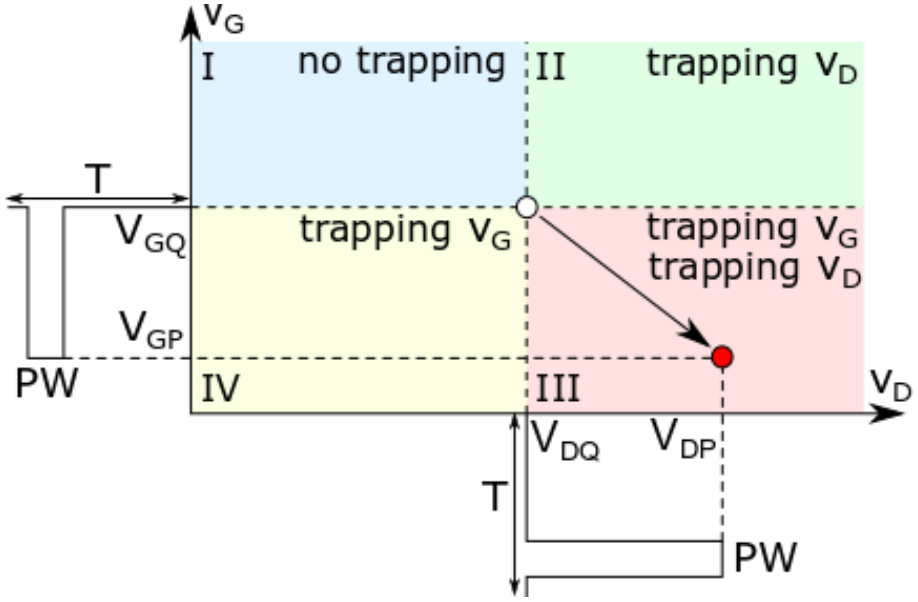


Figure 2.7: Trap activation regions over the voltage domain due to concurrent gate and drain voltage pulses.

Fig. 2.8a shows a practically constant current, meaning that almost no de-trapping has taken place during PW, due to the very short duration of PW w.r.t. the recovery time constants. Finally, Figs. 2.8b-c depict hybrid situations where either gate or drain has induced trapping and vice versa, with a mixed global effect on the resulting transients. Beyond its indirect enhancement of the de-trapping mechanisms, self-heating can have a

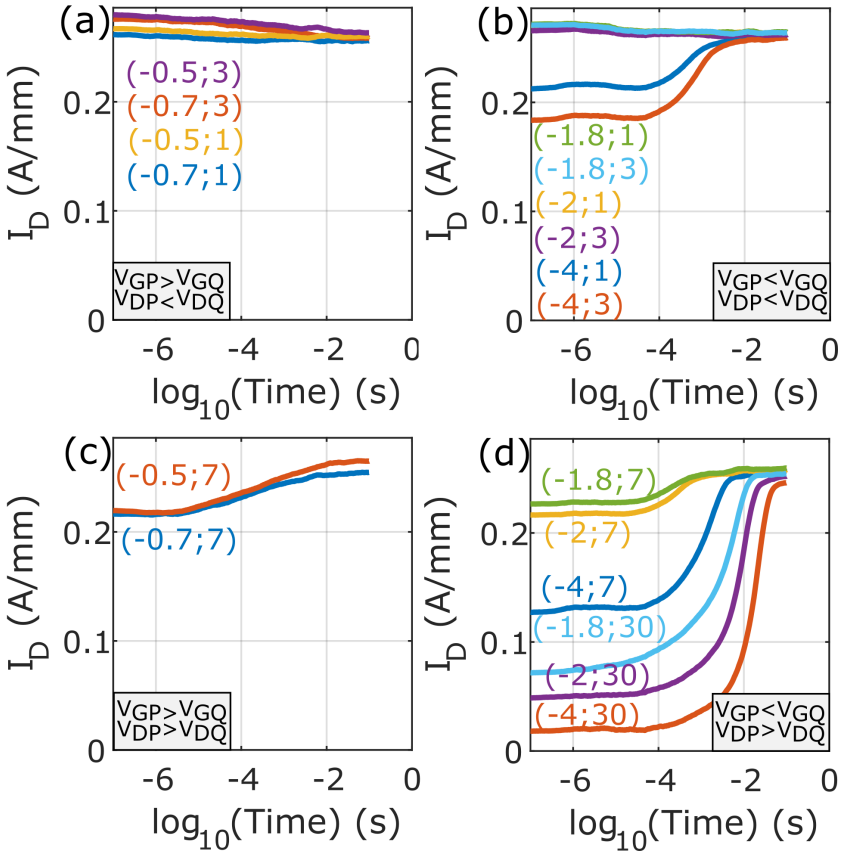


Figure 2.8: Current transients in logarithmic time for all possible reciprocal relationships between pulsed values [PW=500 μs, -4 V $\leq V_{GP} \leq$ −0.5 V, 10 V $\leq V_{DP} \leq$ 30 V, (V_{GP}, V_{DP}) indicated in each plot] and quiescent values (V_{DQ} = 5 V, V_{GQ} = −1.14 V, T_c = 80°C). a) Slight trapping transients, corresponding to a small amount of de-trapping during PW; b) de-trapping transients for V_{GP} = −4, −2 V, corresponding to trapping during PW; slight trapping transient for V_{GP} = −1.8 V, corresponding to limited de-trapping during PW; c) de-trapping transients, corresponding to trapping during PW; d) de-trapping transients, corresponding to trapping during PW.

significant direct impact on the drain current due to the dynamic dissipated power profiles between pulsed and quiescent conditions [159], so that thermal and charge trapping are often hard to separate for a given transient measurement. The thermal sensitivity of the DUT was measured, in static conditions, as being less than 0.6 $\frac{\text{mA}}{\text{mm°C}}$ for all the measured quiescent points for 40 ° $\leq T_c \leq$ 80 °C (T_c being the thermal chuck temperature). Therefore, we can exclude any significant direct thermal effect on the value of the current across a given transient. In addition, the quiescent points considered in the following are chosen to ensure the same quiescent dissipated power across the evaluated cases.

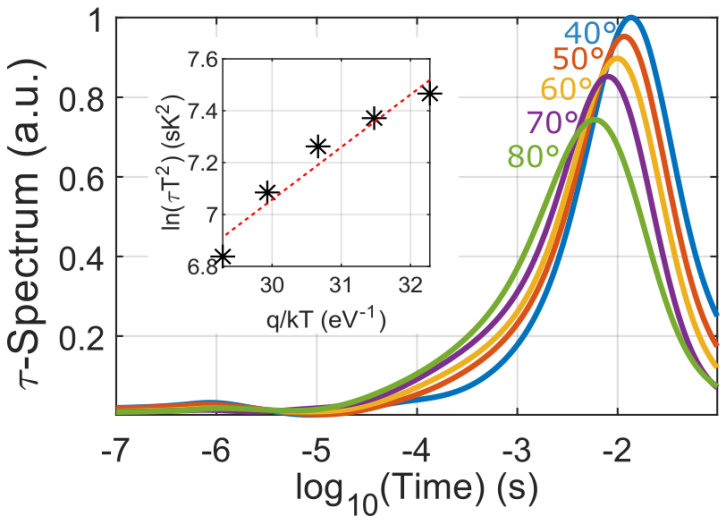


Figure 2.9: Time constant spectra and thermal activation of the trap process. The resulting Arrhenius plot and activation energy are reported. Quiescent Conditions: $V_{DQ} = 5$ V, $V_{GQ} = -1.25$ V. Pulse parameters: PW=100 μs, $V_{GP} = -4$V, $V_{DP} = 20$V.

2.3.3 Arrhenius analysis

The Arrhenius plot for this DUT, extracted with $T_c \in [40, 80]$ °C and compensated for self-heating as in [160] (thermal resistance R_{TH}=110 $\frac{\text{K}}{\text{W mm}}$ as from foundry), displays a single significant trapping process (Fig. 2.9). The trap, with typical time constant in the order of 10^{-2} s, experiences a very weak dependence on temperature and an irregular plot alignment, eventually resulting in an approximated $E_a \simeq 0.2$ eV.
This experimental behavior has already been observed in 150-nm-gate-length AlN/GaN/AlGaN-on-Si HEMTs [161], while being significantly different from AlGaN/GaN-on-SiC HEMTs with similar gate lengths [28], pointing to possible differences in the underlying combined thermal and trapping mechanisms. In fact, since the Si substrate and the AlGaN back-barrier display a significantly higher thermal resistance w.r.t. the GaN buffer-SiC stack of typical HEMTs, the DUT internal temperature is mostly set by self-heating, reducing the sensitivity on T_c and limiting the time-constant variation, thus eventually leading to an ill-conditioned Arrhenius plot estimation. In addition, as will be shown in Secs. 2.3.4 and 2.3.5, for this DUT the emission rate is strongly influenced by the actual quiescent and pulsed conditions. Whereas the Arrhenius characterization would assume a de-trapping of purely thermal origin, strong self-heating and field-induced effects on the carrier emission process, exacerbated by the extremely scaled gate length, depict a rather complex trapping kinetics. In this perspective, here E_a only represents an experimentally-derived *effective* (or *apparent*) activation energy, rather than an intrinsic physical property of the underlying trapping mechanism [47].

2.3.4 Pulsed voltage dependency

We characterize the de-trapping recovery transients due to different voltage amplitudes when both gate and drain excitations cause fast trap activation. In particular, V_{GP} is set to -4 V (deep OFF-state), -2 V (sub-threshold region) and -1.8 V (ON-state), whereas 7 V $\leq$ $V_{DP} \leq$ 30 V. Adopting the definitions used in Fig. 2.1, for each current transient acquired we evaluate the fundamental time-constant and the relative current drop, obtaining Figs. 2.10a and 2.10b, respectively. Notably, τ ranges from less than 100 μs up to more than 10 ms depending on the pulsed voltages, remarking that measuring the dynamic behavior at just one bias, akin to LF Y-parameter, or noise characterizations [28], or at just one pulsed configuration, leads only to a local description that will fall short under LS wideband regimes. The results confirm that the larger the pulsed drain voltage, and the more the gate voltage is pushed towards the OFF-state, the longer the recovery time to quiescent conditions, showing a sub-linear time-constant dependency on the pulsed drain voltage. Moreover, τ decreases as the drain quiescent voltage increases, pointing to a field-assisted emission process [146], as the two quiescent points display the same thermal conditions. The relative current drop shows a similar dependency, exhibiting substantial variations up to a full drop of the quiescent current for the larger drain voltage amplitude. Considering the different trap signatures of this DUT w.r.t. the ones in C- or Fe-doped AlGaN/GaN HEMTs [47] and the use of the AlGaN backbarrier, the presence of buffer dopants is realistically ruled out in determining the characterized trapping behavior. This aspect, together with the presence of the SiN cap layer, which is expected to effectively passivate surface states, suggests that the observed effects are due to defects in the buffer stack, whose impact is augmented by the high electric field induced by the short gate length [131].

2.3.5 Filling pulse dependency

Characterizing the dependency on the filling pulse width provides additional information on the characteristics of the trapping mechanisms [47], [134]. As PW increases, a larger amount of charge gets captured, up to a certain saturation level for the given pulsed voltages [134], [146]. We sweep PW from 50 ns to 500 μs at two different chuck temperatures. Figure 2.11a shows that τ increases from a few ms up to hundred of ms for $V_{GQ} = -1.14$ V, $V_{DQ} = 5$ V, demonstrating that the PW still influences the trapping behavior up to 500 μs. The increase rate of τ is even larger for $V_{GQ} = -1.5$ V, $V_{DQ} = 15$ V, despite showing smaller absolute values from less than 100 μs up to 1 ms. More in detail, both τ and the relative current drop (Fig. 2.11b) show a quasi logarithmic dependency on PW, which is only barely influenced by the chuck temperature. As documented in literature [47], [162], this type of behavior could be ascribed to the presence of point defects, given that the regular reduction of the capture rate would be caused by a localized Coulomb capture barrier, whose height increases with the filling pulse.

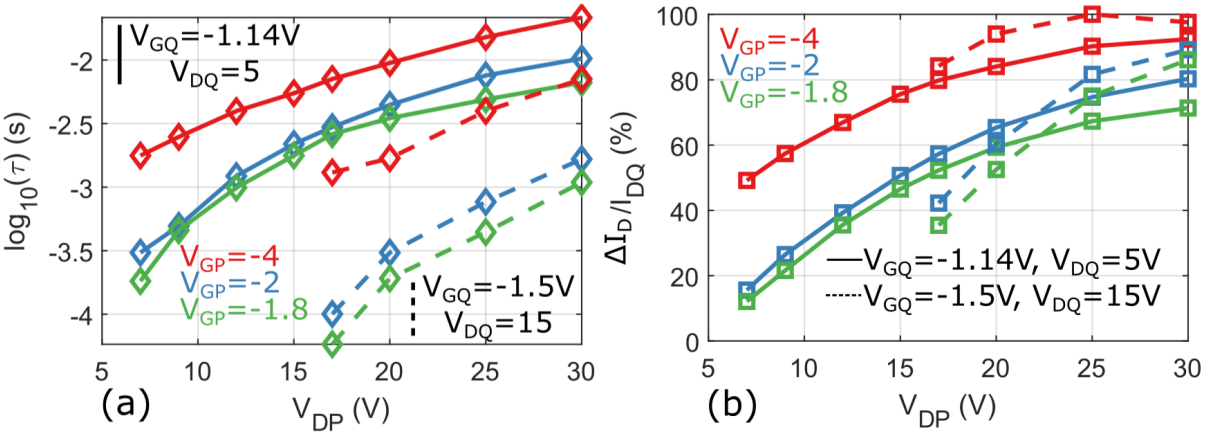


Figure 2.10: Time-constant (a) and normalized current drop (b) dependence on filling pulse gate and drain voltages at a pulse width of 500 μs and chuck temperature $T_c = 80°$C, for two different bias points.

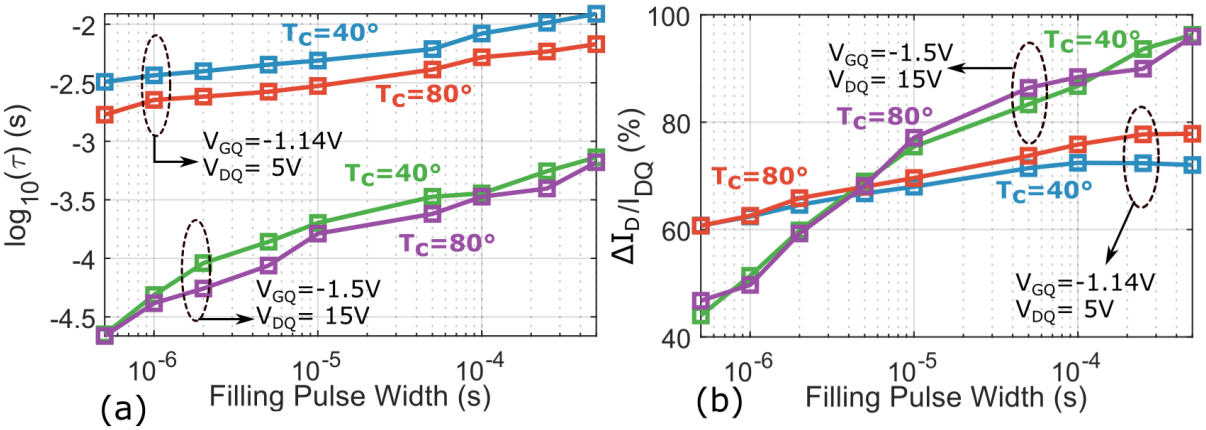


Figure 2.11: Time-constants (a) and normalized current drop (b) dependence on the filling pulse width and temperature for trapping conditions to ensure full current recovery for all pulse widths and temperatures: $V_{GP} = -2$ V, $V_{DP} = 25$ V. Two distinct bias points are reported: $V_{GQ} = -1.14$ V, $V_{DQ} = 5$ V (blue: $T_c = 40°$C, red: $T_c = 80°$C) and $V_{GQ} = -1.5$ V, $V_{DQ} = 15$ V (green: $T_c = 40°$C, purple $T_c = 80°$C).

2.3.6 Analysis of fast trapping transients

The dependency on the filling pulse in the range up to 500 μs indicates that capture mechanisms critically impact the RF performance [28],[155]. Thanks to the fine acquisition of the waveforms (10-ns resolution) and wide-ranging voltage excitation capabilities, we are able to observe the fast drain current transients during the filling pulse, which have been rarely studied in literature [155].

In Fig. 2.12a, we show the trapping transients during PW=500 μs (the longest acquired) with $V_{GQ} = -1.14$ V, $V_{DQ} = 5$ V for different filling-pulse amplitudes. The capture transients critically depend on the applied pulsed voltage, confirming (Sec. 2.3.4) that the electric field distribution strongly influences the trapping mechanisms. Figure 2.12b shows the extracted time constants with the same technique as in Sec. 2.2, with three distinct processes detected: T1 and T3 display a capture-type behavior, while T2 operates as a weak electron-release trapping center. The dominant fast-trapping time constant is the one associated with T1, with a value $\sim$ 300 ns. T2 and T3, characterized by time constants in the μs - 100 μs range, are enhanced by an increase in the pulsed drain voltage. The effects of the three processes and the relative time constants tend to blur for lower values of drain pulsed voltage, leading to a unique broadened spectral peak associated with a strongly non-exponential transient behavior.

2.4 Comparison of trapping dynamics in GaN HEMTs for millimeter-Wave applications

2.4.1 Comparative technology description

Among the many variants of GaN process available, in this section we measure four state-of-the-art processes from four different manufacturers, whose nominal values are

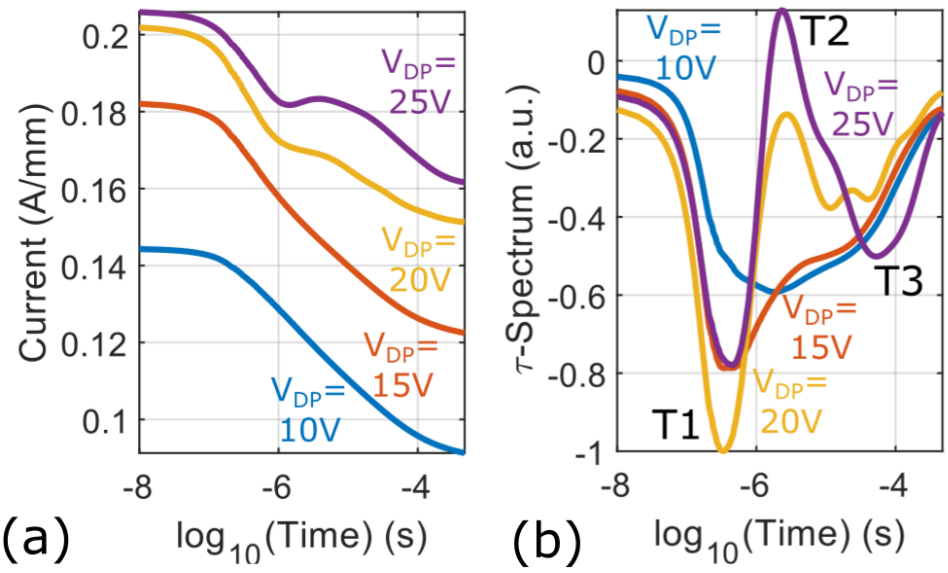


Figure 2.12: Current transients (a) and relative time-constant spectra (b) during trapping pulses for four different pulsed conditions: $V_{GP} = -1.5V$, $V_{DP} = 10-25V$. The bias points is set at $V_{GQ} = -1.14$ V, $V_{DQ} = 5$ V, for $T_c = 40$°C and a filling pulse PW=500 μs.

reported in Table 2.1. In particular, all considered devices have the same number of fingers and periphery ($4 \times 50\ \mu m$). Three devices, labelled as **A**, **B**, and **C**, are 0.15-μm AlGaN/GaN HEMT devices on a SiC substrate. As such, they have superior thermal dissipation properties, displaying thermal resistances in the $\sim 50 - 60$ K/W range.
More in detail, devices **A** and **C** can handle similar pulsed current densities (I_{DSS}) in the order of 1.2 A/mm, and alike breakdown voltages (V_{BD}) in the 60-70 V range. At the same time, **C** shows a much higher nominal f_t and a slightly higher RF power density. Device **B**, instead, shows a much greater breakdown voltage ($V_{BD} > 120$ V), at the expense of a substantially lower current density (0.7 A/mm) and RF power density. Finally, **D** is a different technology employing a 0.1-μm AlN/GaN/AlGaN double-heterostructure on a Si substrate (with a device technology similar to the one reported in Sec 2.3), displaying R_{TH} ∼120 K/W, which corresponds to more than double with respect to the device on SiC substrate. This device configuration can deliver a much higher $f_t > 120$ GHz and similar levels of RF power (at lower base plate temperature) and current densities, yet at the expense of lower V_{BD}. All devices feature peak PAE levels of $\sim 50\%$ and similar gain at millimeter-wave frequencies.

Table 2.1: Details of the measured GaN devices. All DUTs have a $4 \times 50\ \mu m$ periphery.

Label	**A**	**B**	**C**	**D**
Process	AlGaN/GaN	AlGaN/GaN	AlGaN/GaN	AlN/GaN/AlGaN
Substrate	SiC	SiC	SiC	Si
Gate Length (nm)	150	150	150	100
RF Power (W/mm)	3.5	3	4.2 @ 30 GHz	3.3
Gain (dB)	12 @ 30 GHz	13 @ 29 GHz	9 @ 35 GHz	13 @ 35 GHz
PAE (%)	> 45 @ 30 GHz	> 50	> 50 @ 30 GHz	50 @ 35 GHz
PulsedI_{DSS} (A/mm)	1.2	0.7	1.25	1.2
f_t (GHz)	> 35	> 35	> 65	110
R_{TH} (K/W)	54 @ 25°C	56 @ 25°C	61 @ 25°C	128 @ 25°C
V_{BD} (V)	> 70	> 120	> 60	> 36
V_{TH} (V)	-3.5	-1.8	-2.5	-1.5

The static output characteristics at 40 °C for the four devices under test are reported in Fig. 2.13a-b, while the measured voltage thresholds are reported in Table 2.1. Device **A** displays the smallest reduction in dc current with respect to the nominal pulsed values of pulsed I_{DSS}. This fact, together with the absence of a visible self-heating effects, highlights the good thermal performance of this process. Instead, Device **B** and **C** display the largest self-heating effect on the current, with its dc value decreasing for higher V_{DS} values and being significantly different from the pulsed I_{DSS}. Device **D** has been characterized in a restricted SOA due to thermal and voltage ratings of the double hetero-structure of the GaN-on-Si device. Nevertheless, the device displays a significant knee walkout at lower gate voltages, together with a single kink in the knee region. Both signatures have been linked to trapping behavior in GaN HEMTs [136].

2.4.2 Transient multi-bias equal pulse characterization method

The flexibility of the setup in Sec. 2.2.4 allows to fully explore bias voltages across the safe-operating-area of the DUT, hence providing global device performance and complete assessment of the spurious behavior due to dispersive effects. A novel characterization of dynamical effects is performed using two-level pulsed voltage excitations with 50% duty cycle applied concurrently on the gate and drain of the device. The chosen period for the excitation is $T = 100$ ms, which is deemed sufficiently large for practical compact modeling. Indeed, any additional current instabilities with duration of several seconds or more should rather be treated as long-term drifts. The chosen value for T allows for the concurrent measurement from all three channels (gate voltage, drain voltage and current) in a single oscilloscope acquisition at a sufficiently high sampling speed, here selected at 100 MSa/s (10-ns sampling time). The edge-time of the excitation is set to 100 ns as a trade-off between the theoretical requirement for fast transitions and the available setup BW. An example of the employed excitation and generated current transient is shown in Fig. 2.14, showing clean pulsed voltages. In Fig. 2.14, a slow current recovery transient can be observed when the device is pulsed with the voltage transition $(V_{GQ,2}, V_{DQ,2}) = (-2, 36)$ V $\rightarrow (V_{GQ,1}, V_{DQ,1}) = (0, 8)$ V.
The idea behind the use of such waveforms is that, at equilibrium, the internal trapping state of the device is set exclusively by the applied voltages (and possibly, by the baseplate temperature T_c). The equilibrium is then perturbed by the stepped-voltage excitation, and the current transient is measured in order to behaviorally characterize the internal dynamics before a new equilibrium (i.e., the one set by the voltages after the step) is reached.

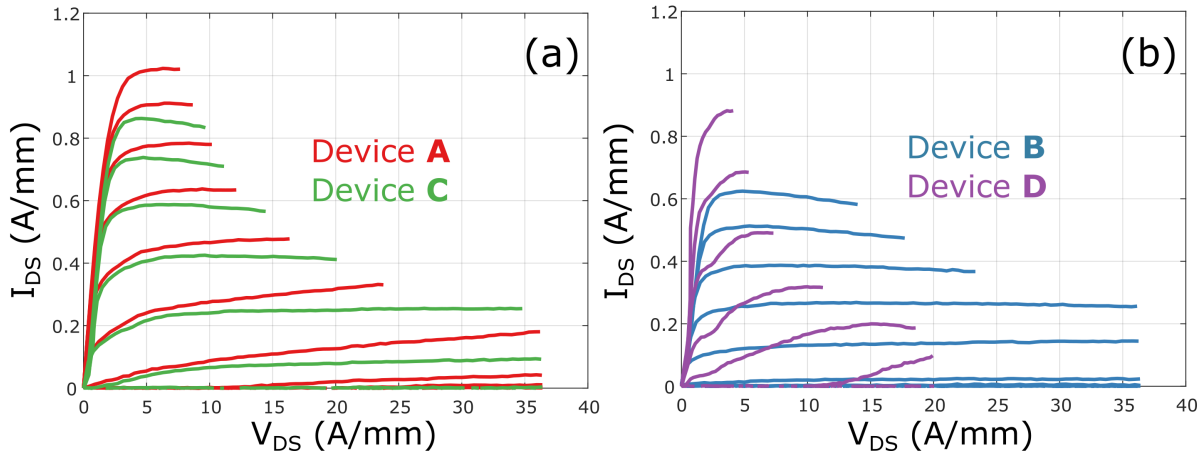


Figure 2.13: Output static characteristic for the four DUTs at 40 °C. a) Device **A** and **C**, V_{GS} swept from -5 V to 0 V in 0.5 V increments. b) Device **B** and **D**, V_{GS} swept from -3.5 V to 0 V in 0.35 V increments.

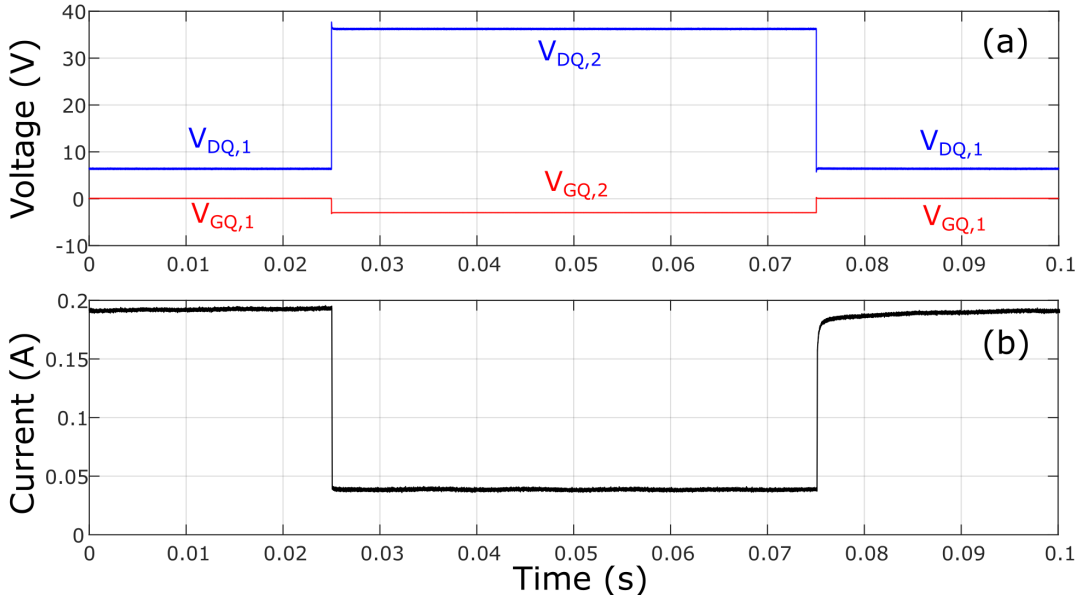


Figure 2.14: Example of (b) current transients acquired for (a) 50% duty-cycle voltage excitation applied at both gate and drain port. Each acquisition allows to evaluate the response to the step-like transition $(V_{GQ,1}, V_{DQ,2}) \rightarrow (V_{GQ,1}, V_{DQ,2})$ and the complementary one $(V_{GQ,2}, V_{DQ,2}) \rightarrow (V_{GQ,1}, V_{DQ,1})$.

In this respect, the drain current is used to probe the inaccessible internal trapping and thermal state of the device. If the period of the pulsed excitation is long enough to allow the current to recover to its dc value, the observed transient will only depend on the gate and drain voltages before and after the pulse (i.e., four variables in total, plus possibly T_c). If N points (V_{GQ}, V_{DQ}) are considered on the gate-drain voltage domain, a global characterization of the overall dynamical behaviour can be obtained by applying pulsed excitations among all the $\frac{N(N+1)}{2}$ possible combinations of two such points, and recording the resulting current transients [101].
Given that the number of measurements increases quadratically with the number of points of interest, a subset of meaningful voltage points have to be selected to provide a reasonably small measurement time and manageable dataset size. This global excitation allows to finely capture and analyze both the slow release and the fast charge capture transients, together with any superimposed dynamic self-heating effects.
The main advantage of this method with respect to existing current-transient characterization techniques [163], [164], including the one in Sec.2.3 lies in the ease of design of experiments. As long as steady-state conditions are reached, no prior hypotheses on the relative value of trapping, de-trapping and thermal time constants has to be made in order to design specifically-tailored multiple filling pulse durations or amplitudes. Therefore, the proposed technique is well-suited for a general-purpose behavioral characterization of

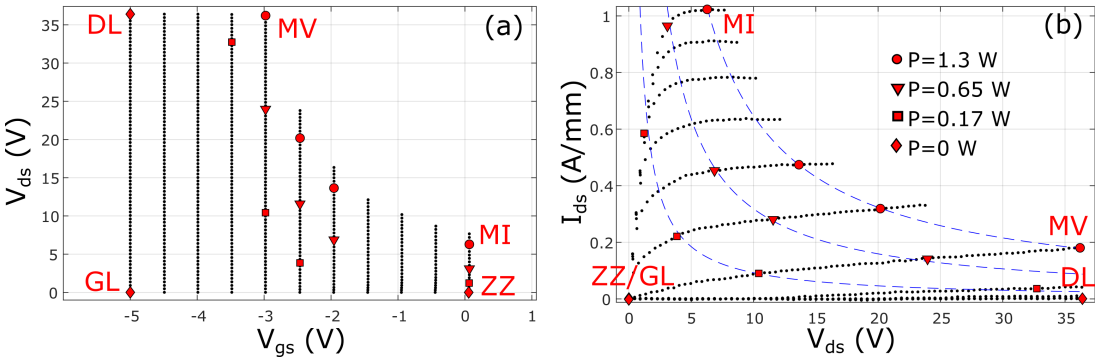


Figure 2.15: Selection of a meaningful subset of quiescent points from the static-IV characteristic of the device. The three key quiescent points for gate/drain-lag characterization *GL*, *DL* and *ZZ* are at zero dissipated power. The other selected quiescent points are found on three iso-thermal IV-characteristics at dissipated powers 1.3 W, 0.65 W, and 0.17 W. (a) V_{GS}-V_{DS} plane. (b) Static-IV characteristic for device **A** at $T_C = 40°$ C.

dynamical effects in GaN devices, whose specific behaviour [101], as it will be discussed in Sec. 2.4.4, strongly depends on the details of the fabrication process and the applied voltages. Other more specific techniques, such as the pulse width variation in Sec. 2.3, might provide more physical insight and complementary information on a specific defect with respect to the proposed 50% pulse method, but cannot provide a technology-agnostic characterization. In any case, many of the well-known characterizations such as pulsed IVs will be available as a subset of the complete dataset.

As an example, let us consider Fig. 2.15, reporting the V_{GS}-V_{DS} plane (Fig. 2.15a) and the corresponding output characteristic (Fig. 2.15b) of the HEMT for the device **A**. Three key voltage coordinates have been selected as from the classical gate/drain-lag characterization, namely: *GL*, corresponding to the minimum gate voltage ($V_{GQ} = -5$ V for device **A**) and $V_{DQ} = 0$ V; *DL*, corresponding to the minimum gate voltage and to the maximum drain voltage ($V_{DQ} = 36$ V as per limitation of the drain driver); *ZZ*, corresponding to (V_{GQ},V_{DQ})=(0,0) V. All these points feature zero power dissipation. Indeed, the measurement of the drain current just after the step-like transitions among these points will allow to quantify the typical gate/drain-lag performance, as well as to monitor the trap-assisted slow current transient from the minimum trapping point *ZZ* to the maximum trapping point *DL*.

Other quiescent points to be included in the transient characterization dataset have been automatically chosen on the dc-IV current characteristic of the HEMT (Fig. 2.15b), which is preliminarily measured. On this characteristic, a limited number of iso-thermal profiles are selected, including the iso-thermal profile with the maximum dissipated power compatible with the maximum baseplate temperature considered for test (here, maximum $T_C = 80°$ C). As an example, for the device **A** in Fig. 2.15b, the maximum-temperature

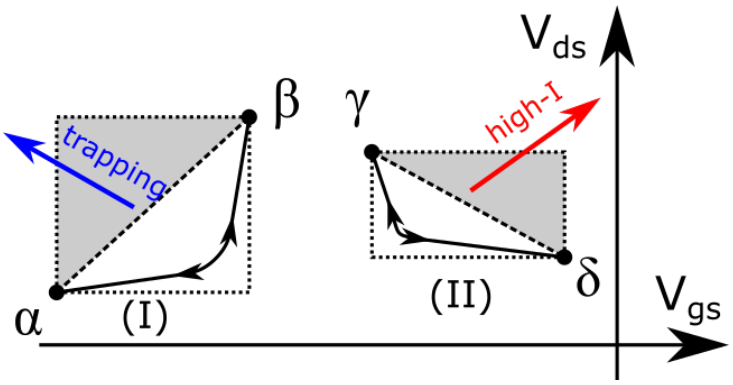


Figure 2.16: Graphical representation of two main cases (I) and (II) for the transitions between two quiescent voltage points. The showed trajectories allow to avoid exciting spurious fast charge capture and non-reversible high-current effects.

iso-thermal profile has been measured at 1.3 W. Then, the maximum-current (MI) and maximum-voltage (MV) points of such iso-thermal profile are included in the subset of quiescent points for transient characterization. Additional quiescent points are finally selected on the other iso-thermal curves by maximizing the voltage swings and the exploration of the V_{GS}-V_{DS} plane. Similarly to [165], this strategy allows to measure current transients subject to the same thermal conditions yet with different applied voltages, allowing for the separation between thermal and trapping dynamic effects. Eventually, the typical dataset collected for each measurement campaign on a given device involved a total of $N = 15$ quiescent voltages, resulting in 120 step-like transitions.
Considering the behavior of traps in GaN and their fast capture mechanisms for increasing V_{DS} and decresing V_{GS} [166], each transition $(V_{GQ,1}, V_{DQ,1}) \rightarrow (V_{GQ,2}, V_{DQ,2})$ and the complementary one $(V_{GQ,2}, V_{DQ,2}) \rightarrow (V_{GQ,1}, V_{DQ,1})$ must be designed to avoid triggering any spurious fast trap capture during the rise/fall step times. Indeed, this unwanted charge capture and its effect on the drain current would contribute to excite a different trapping dynamics that is not unambiguously determined by the user-programmed quiescent voltage pairs, but possibly also by the voltage states rapidly traversed during the transition.
The problematic situations are graphically depicted in Fig. 2.16, showing in grey the regions not to be crossed during the transition. In case (I), the pulse is applied between the α and β voltage points. If the instantaneous voltage, either during to $\alpha \rightarrow \beta$ or the $\beta \rightarrow \alpha$ transition crosses the grey region, the trapped charge might be higher in one of the traversed states than in either α or β. As charge capture is typically orders of magnitude faster than charge release, even a fast transition in the forbidden area would be enough to set a given trapping state [152]. Then, the observed transient would be dominated by the slow de-trapping from this intermediate spurious state, obscuring the actual $\alpha \leftrightarrow \beta$ transients of interest. Conversely, by following the drawn trajectory in Fig.

2.16, any unwanted side effect is avoided. In case (I), considering that the user-selected β point is the one with the highest current of the whole trajectory, there is no concern that the device will dynamically exit the SOA.
In case (II), γ is expected to be the point with the highest amount of trapped charge, whereas δ is the one with the lowest amount, irrespective of the chosen trajectory. Therefore, there is no concern of spurious intermediate trap captures impacting the transient response. On the other hand, a trajectory crossing the grey area would possibly display increased stress conditions with respect to either δ and γ, thereby exiting the SOA of the device. Even if this crossing happens during the fast transition between the voltage points, the intermediate points might feature higher instantaneous currents and junction temperatures, possibly causing damage to the device or altering the observed transient response. In all cases, the required trajectory can be enforced by accurate timing of the pulsed waveforms, with a carefully controlled overlapping of the 100-ns-long pulse edges.

2.4.3 Gate/drain lag

In order to compare the dynamical effects of the different device technologies introduced in Sec. 2.4.1, a significant subset of the overall set in Fig.2.15 has been selected. The set contains the *GL*, *DL*, *ZZ*, *MV* and *MI* points in order to provide a global characterization of the thermal and trapping phenomena. The actual voltage and dissipation values used on the iso-thermal line for *MI* and *MV* that were examined for each device technology are described in Table 2.2.

Table 2.2: Voltage values (V_{GQ}, V_{DQ}) and power dissipation points selected for the devices **A**-**D**.

Label	**A**	**B**	**C**	**D**
Gate Lag Point - *GL* (V)	(-5, 0)	(-3.5, 0)	(-5, 0)	(-3, 0)
Drain Lag Point - *DL* (V)	(-5, 36)	(-3.5, 36)	(-5, 36)	(-3, 20)
Max Voltage Point - *MV* (V)	(-3, 36)	(-1.4, 29)	(-2, 17.5)	(-1.2, 18.5)
Max Current Point - *MI* (V)	(0, 6.3)	(0, 6.8)	(0, 4.6)	(0, 3.9)
Power Dissipation (W)	1.3	0.8	0.8	0.7

Figures 2.17a-b detail all the possible transitions among these five points displaying a non-zero current, as its value is used to probe the trapping state of the device. Each transition displays a different combination of capture/release and thermal transients, which is often overlooked in typical gate/drain-lag characterizations. In particular, the *MV*-*MI* transitions can be used to separate self-heating effects and charge trapping ones, given that the assumption of an instantaneous self-heating suggested in [149] might not be realistic for all device structures.

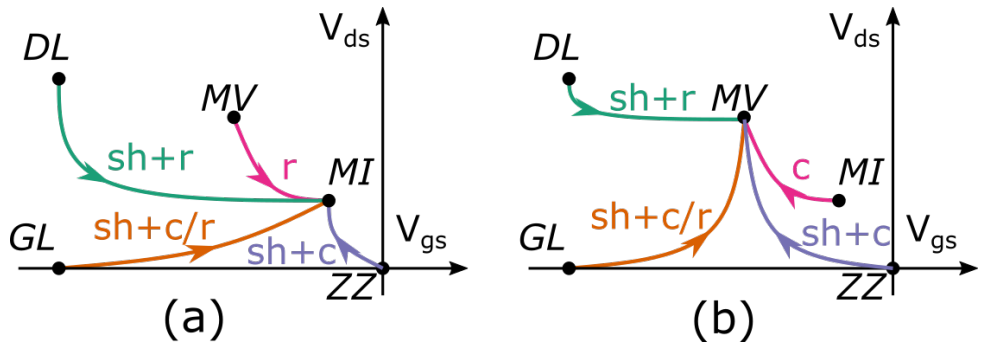


Figure 2.17: Graphical representation depicting all the relevant trapping/de-trapping transitions on the $V_{GS} - V_{DS}$ plane from the *ZZ*, *GL*, *DL* points to the *MI* (a) and *MV* points (b). The involved processes are highlighted on the transitions: self-heating (*sh*), charge capture (*c*) and release (*r*).

This distinction can be observed, for example, in Fig. 2.18, which details all the transitions towards the *MI* point for device **D**. While the charge release and self heating transient from *DL* displays a temporary decrease of current at the ∼1 ms mark, the same signature is absent in the charge release iso-thermal transient from *MV*, despite the very similar voltage pulsing conditions. This difference in behavior in this case can therefore likely be attributed to the existence of a dynamic self-heating phenomenon. Alternatively, current drops can also be interpreted as pointing to the spurious charge capture processes, which have indeed been shown to be present in some GaN technologies [47], but the use of iso-thermal transients seems to exclude this hypothesis in the DUT. The use of a complete dataset can therefore be employed in order to partially untangle the reciprocal relationship between thermal and trapping phenomena, avoiding spurious identifications of each dynamical effect.
As an example of the proposed global characterization, time-domain transients for device **A** at $T_c = 40°C$ are reported in Fig. 2.19. In all the cases, the current displays almost complete recovery to equilibrium, identified by the common endpoint of all the transients. By first examining the transients towards the *MI* point (Fig. 2.19a), it is possible to notice that the measurement from *ZZ* shows a minor trapping and self-heating current decrease transient. Similar behaviour is observed when pulsing from the *GL* point, highlighting the relative absence of strong gate-lag effects in the device. On the other hand, relevant increasing current transients are excited when pulsing from either *MV* or *DL*, pointing to the existence of significant slow de-trapping dynamics triggered by decreases in drain voltage. While the initial current drop with respect to the equilibrium dc value is similar in both cases, the transient displays a significantly different emission time. The difference can be likely attributed to the different initial thermal state between the two cases, as the *MV* point is iso-thermal with *MI*, while *DL* is not. The difference in V_{GS} between these two cases is instead likely to have a minor impact on the difference in transients, as

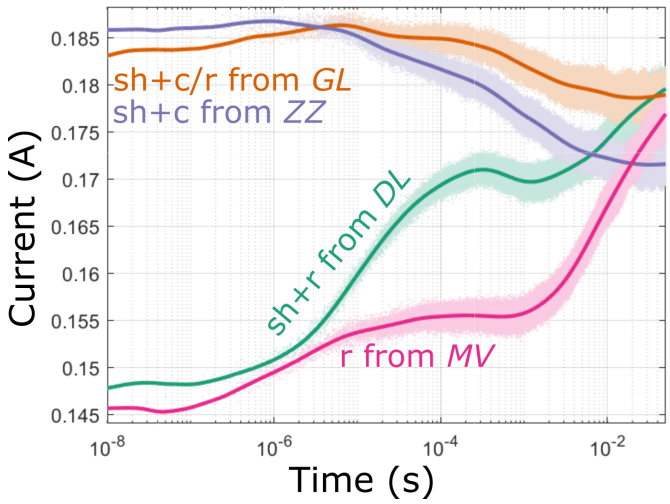


Figure 2.18: Time-domain current transients from *ZZ*, *GL*, *DL* to *MI* points for device **D** at $T_c = 40°C$. Lighter shades of color identify acquired points, while solid lines are smoothed versions of the current transient.

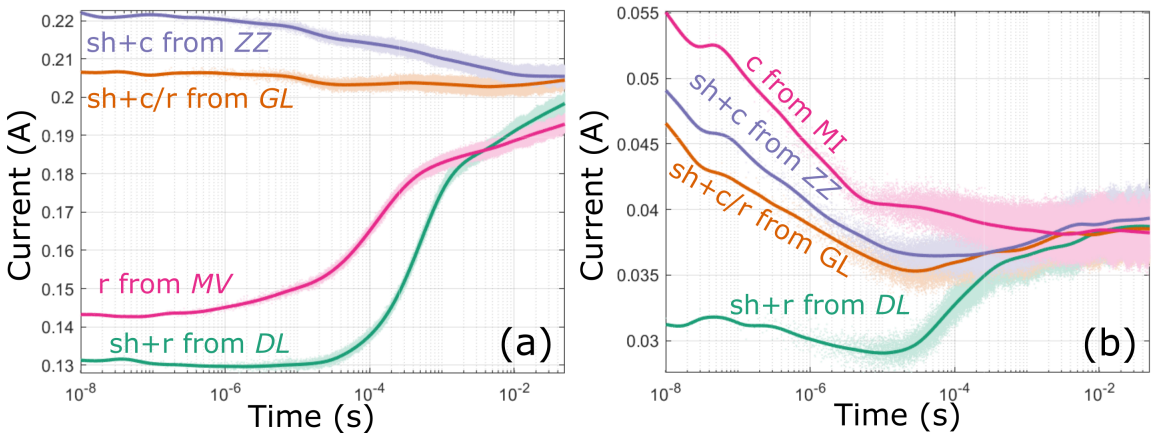


Figure 2.19: Time-domain current transients from *ZZ*, *GL*, *DL* to (a) *MI* and (b) *MV* points for device **A** at $T_c = 40°C$. Lighter shades of color identify acquired points, while solid lines are smoothed versions of the current transient.

shown by the limited gate-lag transient.
The observation of the transients towards the *MV* point (Fig. 2.19b), highlights the existence of a fast (∼100 ns) capture process when observing the currents from the isothermal *MI* point. Similar behaviour is observed also in the pulses from the *ZZ* and *GL* point, highlighting that the observed transient is mostly due to the trapping state in the

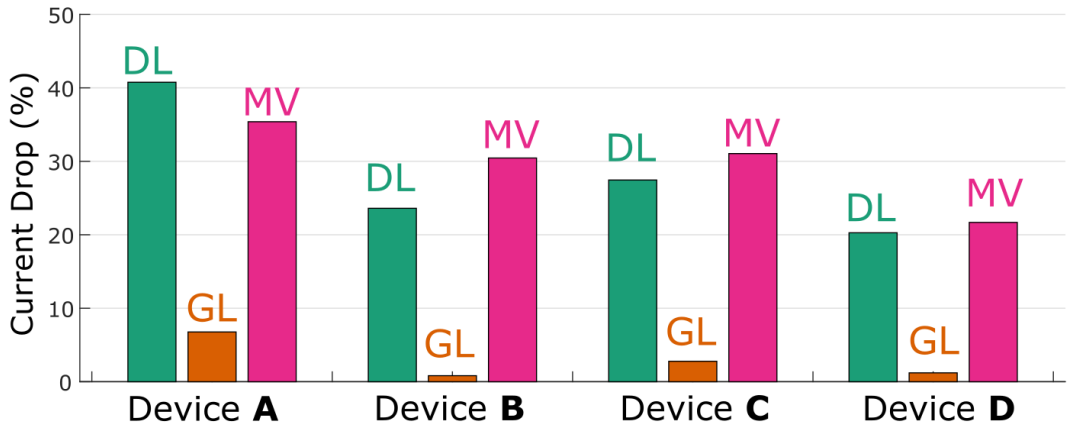


Figure 2.20: Percentage current drop measured at the *MI* point by pulsing from *GL*, *DL* and *MV* points with respect to the value obtained from the *ZZ* point for devices **A**-**D** at $T_c = 40°C$.

high drain voltage *MV* point, with minor influence of self heating. On the other hand, the drain-lag case displays an observable de-trapping transient, marked by the overall increase of the current towards equilibrium. The transient also embeds a dynamical self heating effect, identified by the temporary decrease of the measured current in the 1-10 μs range. The co-existence of both phenomena, as highlighted in the previous section, has to be taken into account for the formulation of behavioral models, as well as for the investigation of specific physical mechanisms. Similar considerations can be drawn for devices **B**, **C**, and **D** (not shown).

The collected dataset can be also used in order to extract gate-lag and drain-lag metrics of each device, allowing for a detailed technology comparison. By examining the start of each transient (in this case, the first 100 ns), one deduces the current that would be measured within the short pulse in a classical PIV setup. Figure 2.20 compares the percentage current drop measured at the *MI* point by pulsing from DL, GL and *MV* points with respect to the value obtained from the *ZZ*. Despite the different technologies, all the devices display a reduced gate lag effect. Gate lag has been linked mostly to surface states, whose effects have been shown to be greatly reduced by surface passivation and the use of field plates [57]. On the other hand, drain lag effects are still quite relevant in all the cases, with current drops in the order of 20-40%. The drop observed for the iso-thermal *MV* point closely mimics the one seen for the similar *GL* point even if, as pointed out in the transient analysis, the dynamic behaviour can still be significantly different.

2.4.4 I-DLTS and Arrhenius analysis

As described in Sec. 2.2.3, an Arrhenius analysis is normally used to identify trap energy levels. The actual current levels and trap signatures strongly depend on the process technology and applied voltages, and can give valuable information on the actual transport nature and locations of the trap centers. In order to compare the dynamic behaviour of the different technological solutions adopted in devices **A**-**D**, de-trapping current transients were examined after the $DL \rightarrow MI$ transition at different baseplate temperatures in the range $T_c = 40 - 80°$ C. In this way, it is possible to activate buffer-related traps at high drain voltages, which is the worst-case scenario in terms of trapping, provided that gate-related effects have been observed to be quite limited for all the examined devices.
Trapping time-constants and transient amplitudes specific to each de-trapping process can be found from the peaks in the I-DLTS trace, as outlined in Sec.2.2. As the different devices display different current densities, the I-DLTS in this case is normalized by the dc current at MI, in order to assess the impact of the trapping transient relative to the respective quiescent value. Both the current transients and the I-DLTS traces are reported for all devices in Figs. 2.21a-d. Figure 2.22 collects the resulting Arrhenius plots, together with the apparent activation energy for the traps in each device.
As it can be observed, all devices under test display some degree of trapping, albeit with different severity and characteristics. Devices **A** and **B** display a single pronounced peak in the 100 μs - 1 ms range, with an activation energy of 0.55 eV and 0.59 eV, respectively. The measured energy levels and the single-peaked signature have been linked in literature with Fe doping in the GaN buffer layer [47]. The intentional introduction of acceptors (such as C or Fe) is typically used to enhance the performance of the buffer, by increasing the substrate resistivity and the breakdown voltage while, at the same time, improving electron confinement and reducing short-channel effects.
Devices **C** and **D**, instead, present strong non-exponential behaviour, with significantly broadened spectral peaks in the I-DLTS traces and a less pronounced current transient. In both cases, the trapping mechanisms still show a significant thermal activation of the de-trapping process, ruling out temperature-independent transport phenomena [47]. The linear fit predicted by the Arrhenius model is also less accurate in both the examined cases. These peculiar signatures are likely to rule out acceptor buffer doping and instead can be be traced back to a large variety of defects introduced by the fabrication process [47], whose exact identification lies beyond the scope of the present paper. In particular, the GaN-on-Si device (**D**) operates at a greater junction temperature due to higher thermal resistance, and it presents two different trapping locations (one at $\sim$ 10 μs and the other at $\sim$ 10ms) with significantly different activation energies. Regarding this device, the presence of an AlGaN back barrier for improving confinement realistically rules out the presence of intentional buffer doping. However, the exact cause is yet subject of investigation.
In all the examined cases, due to the conditions chosen for the I-DLTS characterization,

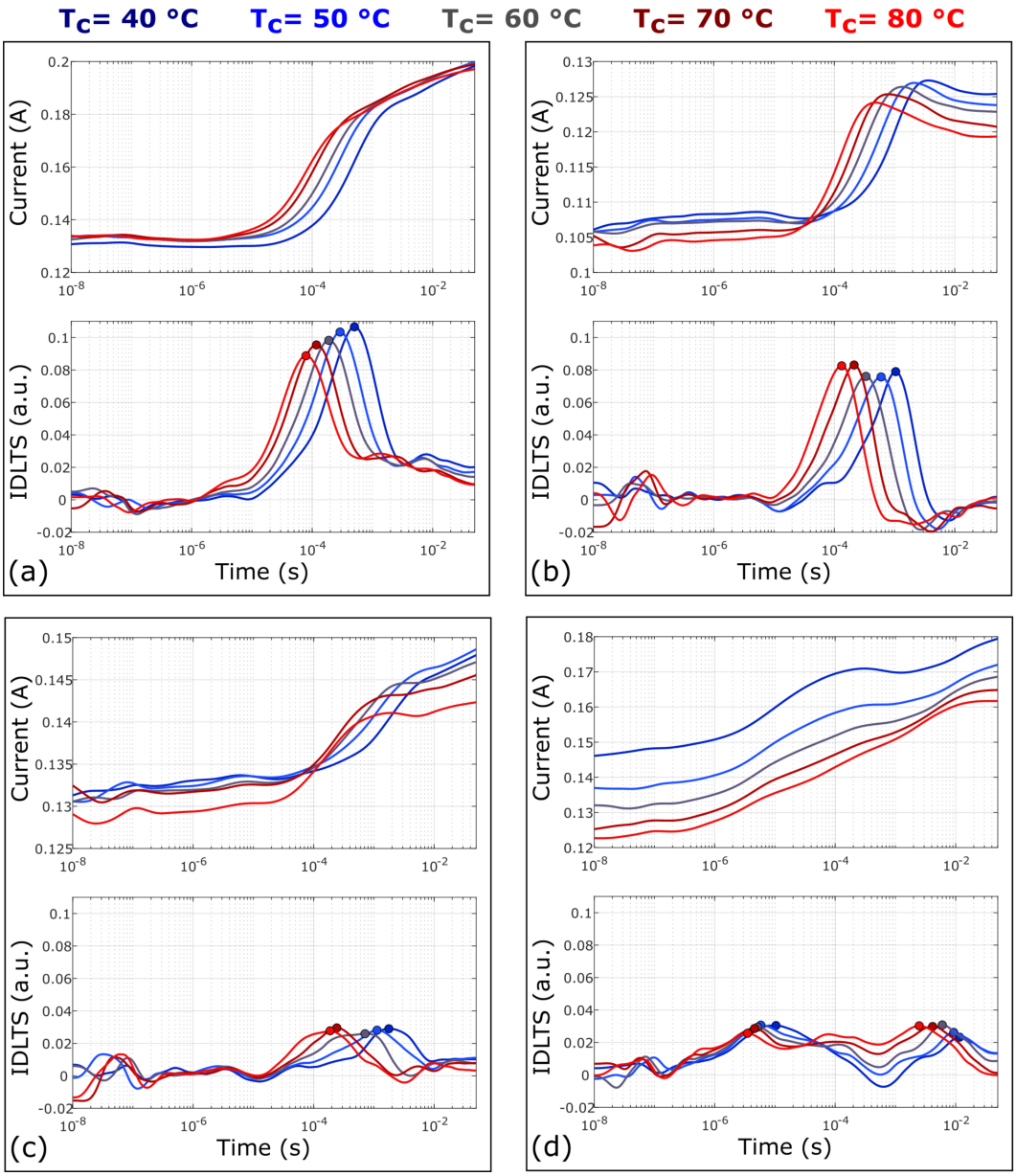


Figure 2.21: Comparison of the current transients (top plots) and I-DLTS traces (bottom plots) among the four studied technologies (devices **A**-**D** in (a)-(d) respectively) for a baseplate temperature range $T_c = 40 - 80°C$. The transients are taken from the *DL* point to *MI* point for each technology, as from Table 2.2.

the transient is not iso-thermal and is indeed subject to thermal dynamics. This effect can be noticed, for example, in Fig. 2.21b and d, where the increase of current due to de-trapping is followed by a decrease due to self-heating. The comparison with the iso-

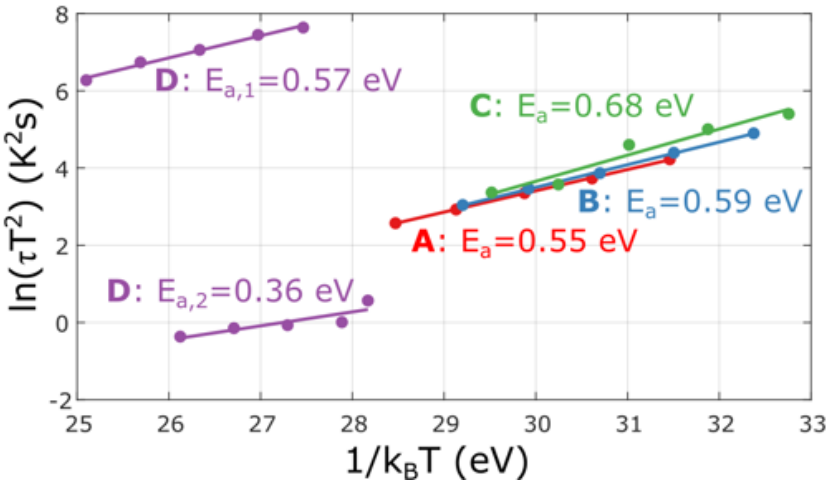


Figure 2.22: Comparison of the Arrhenius plots among the four technologies considered. SiC-based devices (**A** to **C**) share a similar trapping signature and apparent activation energy, whereas the Si-based device (**D**) reveals a different behavior involving two trapping signatures.

thermal transients discussed in Sec. 2.4.3 allows to reasonably exclude the existence of specific charge capture processes in the examined situations.

2.5 Conclusions

In this chapter, pulsed current transient techniques for the identification of trapping effects have been introduced and reviewed. An updated low-frequency measurement bench has been presented, and the proposed methods were used to characterize the charge and release dynamical phenomena in several state of the art sub-0.15 μm GaN devices for mm-wave applications, both on SiC and Si substrates.
In particular, the detailed analysis of a 100-nm double-heterojunction AlN/GaN/AlGaN-on-Si RF HEMT process has revealed a fundamental de-trapping time-constant in the 0.1-1 ms range, showing reduced thermal sensitivity and strong field dependency. Differently from typical AlGaN/GaN HEMTs, this behavior is not attributable to intentional doping, but indicates trapping in the buffer stack, whose impact is exacerbated by the high electric field induced by the short gate length. In addition, the logarithmic dependency on the filling pulse width suggests the presence of point defects. The capture transient analysis, performed for the first time in this technology, has revealed a much smaller time constant at $\sim$ 300 ns, and further effects in the μs range.
Moreover, a novel time-domain charge trapping characterization technique for GaN HEMTs has been proposed. The method generalizes several well-known trap signature character-

ization techniques, such as gate/drain-lag and I-DLTS, and expands their capabilities by providing a complete exploration of the transient behavior across the SOA of the DUT. The acquired dataset is used in order to compare four different technologies.
All the examined technologies display different degrees of charge trapping, with extracted time constants on the order of 10 μs - 10 ms, which can critically interfere with the performance of typical telecommunication applications with modulation frequencies in the same range. While gate-lag phenomena related to surface traps seems modest with respect to previous technological generations, drain lag is still significant for all the examined devices, with the greatest effect displayed by technologies employing buffer acceptor doping for charge confinement.
In conclusion, the results show a large variety of charge capture/release behaviors among different processes. Several underlying physical mechanisms and interactions with self-heating phenomena have been observed, some of which are traceable to the use of short gate length devices and novel process stacks. Single-bias or traditional PIV characterization methods have been shown to, in the general case, poorly represent the global trap behavior under LS conditions. Moreover, the choice of excitation strategy and timings for these behavioral characterizations has to take into account the underlying trapping kinetics, which can be characterized using I-DLTS and related methods. These findings indicate the need for the formulation of novel general-purpose LS trap models, as many of the existing approaches give just a partial description of the trapping dynamics in many modern GaN HEMT technologies. These extensions would then be fed as additional components into existing behavioral and circuit-level modeling flows, improving the predictive performance of models used in PA design.

Chapter 3

Experimental radio-frequency characterization of GaN devices and PAs

3.1 Introduction

As discussed in Chapter 2, dispersive effects due to charge trapping dynamics have time constants in the μs-s range, and their impact is best studied in terms of the influence on the IV characteristics of the device. These time constants intervene at frequencies much lower than those adopted in typical RF standards. Indeed, for a CW waveform in the GHz range, the fast periodic variation of the applied voltages does not allow the slow detrapping process to take place, leading to a fixed trap occupation level that just depends on the peak amplitude value of the applied signal.
Nevertheless, trapping has been shown to have a great impact in many RF PA and RF switch applications [167]–[169], due to its interaction with amplitude-modulated signals. Under this type of excitations, the peak voltages seen by the DUT vary with a modulation frequency which can be comparable to the observed detrapping time constants, leading to several dynamic effects that might impair PA operation. In pulsed PAs (e.g., radar), the performance depends on the duty cycle and pulse repetition frequency [170]. In wideband telecom transmitters, the interactions with trapping might prevent effective linearization of the PA [170], given that the 5G standard will feature high PAPR signals with long frame durations (10 ms) and narrow sub-carrier spacings (down to 15 kHz).
While LF characterizations provide important information on the physical phenomena that underlie trapping, they have been reported to provide inaccurate estimates on the severity of the RF effects in some cases [171]. Therefore, for the purpose of GaN RF PA and device characterization, the impact of trapping has to be directly measured in terms of the variation in the observed high-frequency performance [59], [155]. This requires the

use of different characterization techniques that employ modulated RF signals, which can closely mimic the typical operating conditions seen in application.
Several of such methods have been investigated during the PhD research and the experimental results are reported in this chapter. Section 3.2 describes the work published in [172], where the low-frequency double pulsed IV characterization of a 100-nm GaN-on-Si DHEMT is complemented with the concurrent measurement of S-parameters during the pulse. Section 3.3 is based on [69], which reports the dynamical trapping behavior of a similar device under large-signal RF operating conditions, using different pulsed and two-tone excitations. Section 3.4 reports the results in [173], which proposes a novel EVM measurement technique to characterize the effect of trapping and other long memory phenomena directly at PA level using realistic modulated excitations. Finally, conclusions on the impact of trapping dynamics on the RF performance of electron devices and PAs are drawn in Sec. 3.5.

3.2 Narrow-pulse-width double-pulsed S-parameters measurements of 100-nm GaN-on-Si HEMTs

3.2.1 Characterization technique

Multi-bias S-parameters datasets are used in the identification of several empirical large-signal models for RF electron devices [174]. However, similarly to static IV characterizations, S-parameters measured at different fixed dc biases will generally display different internal trapping states, leading to the extraction of non-isodynamic differential parameters for the HEMT of interest. The use of these parameters might lead to non charge-conservative behavior of the overall device model [175], which is typically an undesired nuisance.
On the other hand, as discussed in Sec. 2.2, the use of double pulsed excitations enables the setting of constant trap occupation state across different gate and drain voltage conditions. The resulting DPIV characteristics have been shown to provide a good prediction of the HEMT behavior. Therefore, the use of double pulsed S-parameters has been proposed [175] as an useful tool to improve large signal RF modeling performance of GaN HEMT. At the same time, the comparison between static, pulsed and double pulsed S-parameters can give an indication of the impact of trapping on the RF characteristics of the device. While current degradation mechanisms are well-known and readily observed from the LF IV characteristics, the possible impact of trapped charge on the equivalent capacitances (or equivalent charge functions [87]) has to be evaluated at high-frequency. Indeed, the relevance of trapping on the fast charge storage phenomena in GaN HEMTs has been less studied, and no widespread consensus exists on the possible signatures [174], [175].
In this section, we enhance the capabilities of the low-frequency setup in [107] by enabling the measurement of (double-)pulsed S-parameters [176]. Differently from similar

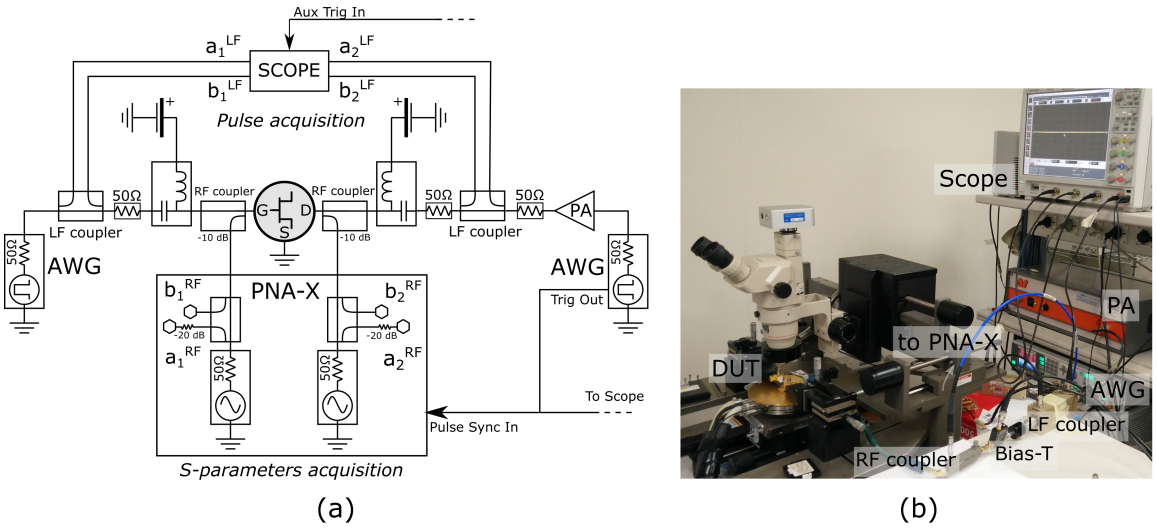


Figure 3.1: Combined LF-RF measurement setup for pulsed S-parameters measurements. (a) Block diagram. (b). Photo of the setup.

setups [175], the implemented measurement bench allows for the excitation of fast pulses, achieving the acquisition of pulsed S-parameters within very narrow pulse widths, down to hundreds of ns.

3.2.2 Measurement setup

The device-under-test (DUT) is measured on-wafer by means of a thermally-controlled probe station. Similarly to other published measurement benches [108],[177], the developed setup (shown in Fig. 3.1), is composed by an LF part, here used for the pulsed baseband excitations, and an RF part, for S-parameters measurements.
Differently from other pulsed systems, such as the one introduced in Sec. 2.2.4, the LF pulses are generated by a two-channel Arbitrary Waveform Generator (AWG) with 50 Ω output impedance, sampled by bridge couplers, and applied through the ac path of the bias-tees. Since all components are 50-Ω-matched (possibly, 50-Ω attenuators can be inserted to improve the VSWR), the setup guarantees wide frequency range, as the BW is typically limited by the lowest operating frequency of the bias-tees (10 kHz in this setup) and by the minimum among the highest operating frequency of the couplers (1 GHz), of the AWG (120 MHz) and of the load bench amplifier (250 MHz). Hence, sharp pulses with just a few ns raise/fall times and down to 50 ns pulse widths can be obtained without any pulse pre-emphasis. This implies a key advantage, considering that GaN trap behavior is especially dependent on the peak voltages applied. Since the controlling variables are the incident power waves generated by the AWG, an iterative algorithm allows to converge at targeted pulsed voltage values on unmatched DUTs. The usage

of a LF network analyzer configuration, where four channels of an oscilloscope are used as receivers, allows for applying well known calibration algorithms such as short-open-load-thru (SOLT) to compensate for impedance mismatches, as well as amplitude/phase calibration up to the on-wafer probe tips.
Concerning the RF part, the acquisition of the S-parameters is achieved by means of a 10 MHz - 26.5 GHz Keysight PNA-X Network Analyzer and its internal couplers. The combination of RF and LF waves is obtained with two additional RF directional couplers (one per port) featuring 10-dB coupling starting at 1 GHz. Considering that the pulse spectra are confined well below 1 GHz, the coupled port can be safely connected to the PNA-X ports without the risk of saturating or damaging the PNA-X receivers. At the same time, the RF coupler thru connection is practically transparent for the pulsed waveforms ($<$ 1 dB insertion loss compensated by the LF SOLT calibration). The RF waves are therefore applied through the 10-dB coupled arms, and external attenuation (20 dB) was added to the paths of the RF incident waves in order to have the same power level at all PNA-X receivers. For a better separation of the two paths, improved RF impedance termination, and to increase the receivers' dynamic range, a suitable diplexer can be used instead.
Both the scope as well as the the PNA-X receivers are triggered by a *Trigger Output* signal generated by the AWG, so that the S-parameters measurements can be performed as a point-in-pulse acquisition in a synchronized way. By varying the mutual trigger delay, pulse profiling can also be obtained. Using the PNA-X wideband acquisition mode [176], the maximum PNA-X Intermediate Frequency (IF) BW is 15 MHz, practically limiting the time domain resolution of pulse profiling to 50 ns. This, in turn, sets a lower limit for the pulse widths, considering that a sufficient number of S-parameter samples must fall within the pulse. In the following, we have chosen pulse width $\tau = 500$ ns and period $T = 50$ μs ($\rho = \frac{\tau}{T} = 1\%$).
As shown in Fig. 3.2 PIV and DPIV stimulus waveforms are concurrently applied at both gate and drain ports. The S-parameters and current measurements are obtained by averaging in time the acquisition points gathered during the 500-ns measurement pulse. For the DPIV case, the pre-pulses feature a duration of 500 ns and are configured to generate the minimum $V_{GS} = -2$ V and the maximum $V_{DS} = 24$ V, and therefore exciting the maximum amount of dynamically trapped charge.
The technology under evaluation is 100nm-gate-length GaN-on-Si introduced in Sec. 2.3. The DUT measured in this work is a 2x70 μm device in common-source configuration biased in deep class AB at $V_{DS,Q} = 12$ V and $I_{DS,Q}$=5 mA.

3.2.3 Pulsed IV characteristics

First, the effect of traps on the behavior of the DUT has been quantified by comparing the static IV versus the single- and double-pulsed IVs, as shown in Fig. 3.3. For both pulsed characterizations, the quiescent point is taken to be $V_{DS,Q} = 12$ V, $I_{DS,Q}$=5 mA,

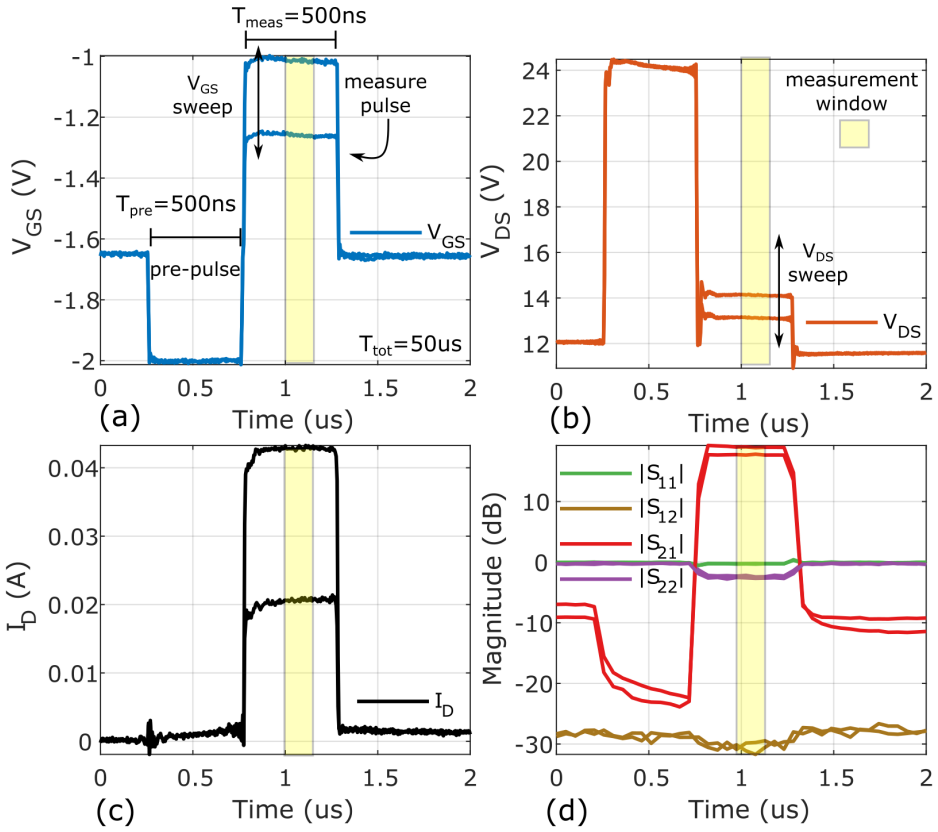


Figure 3.2: Time-domain evolution of double-pulsed measurement (pre-pulse: $V_{GS} = -2$ V, $V_{DS} = 24$ V) for different measurement pulses at $V_{GS} = -1.25, -1$ V and $V_{DS} = 13, 14$ V. (a) Gate voltage. (b) Drain Voltage. (c) Drain Current. (d) S-parameters.

emulating the conditions of a deep class-AB bias point. For the static IV, on top of the self-heating effect, each point of the characteristic belongs to a different charge-trapping state: the higher the V_{DS} and the lower the V_{GS}, the larger will be the amount of trapped charges.

As discussed in Sec. 2.2, the DPIV characteristics are the most suitable ones to describe typical large-signal operation in RF PAs, (e.g., class-AB/E/F) [140], [170]. Notably, for this DUT they show the largest current drop (∼20 % vs the static current) and a clear increase of the knee voltage. These results show compatible trapping signatures with respect to the PIVs from different quiescent points reported in Sec. 2.3, for a device using the same technology albeit with a different periphery.

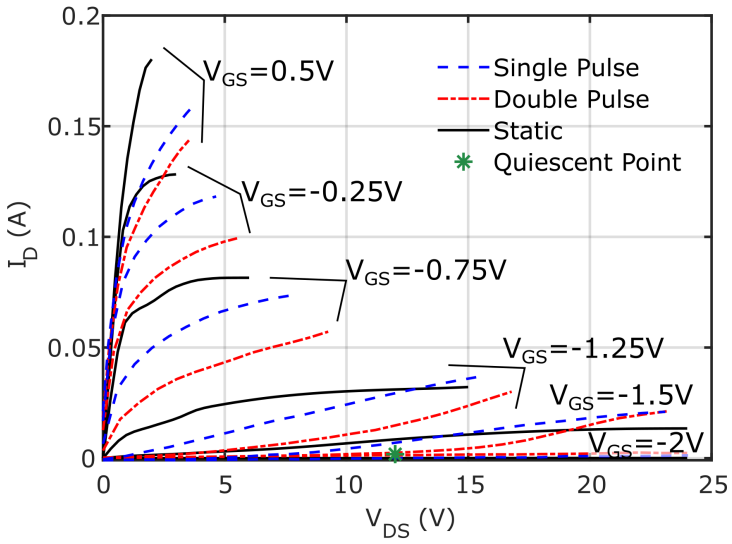


Figure 3.3: Single-pulsed, double-pulsed with pre-pulse $V_{GS} = -2$ V, $V_{DS} = 24$ V (all pulsed from quiescent point $V_{DS,Q} = 12$ V, $I_{DS,Q}$=5 mA) and static IV output characteristics for the 100-nm 2x70 μm OMMIC device at chuck temperature $T_C = 40°$ C.

3.2.4 Pulsed S-parameters

The S-parameters under static, single-, and double-pulsed conditions have been acquired in the 1 GHz - 7.5 GHz frequency range, and they are compared in Figs. 3.4-3.6. Figure 3.4 shows the small-signal gain ($|S_{21}|$) and transconductance ($g_m = \Re\{Y_{21}\}$) at 5 GHz over the transistor output characteristic and transcharacteristic, respectively, in the case of static, single-pulsed and double-pulsed regimes. Coherently with the PIV characteristics, whose derivatives with respect to the voltages are the LF limit values of S-parameters, it can be seen that the single- and double-pulsed cases mostly show lower gain and transconducatance values with respect to the static case.
More in detail, Fig. 3.5 reports the difference in linear units between the small-signal gain in case of single-pulsed and double-pulsed regimes ($|S_{21}|_{SP} - |S_{21}|_{DP}$), highlighting the effect of the different trapping conditions. In fact, the double-pulsed values are substantially smaller than the single-pulsed ones in most parts of the voltage plane. In other regions, the values are similar or even slightly higher for the double pulsed case.
The Smith chart in Fig. 3.6, reports the behavior across frequency (1-7.5 GHz range) for the S-parameters measured in different conditions for two representative voltage points (A and B in Fig. 3.5). As expected from LF characterizations, the effect is more marked on the S_{21} and S_{22} parameters, while being negligible in all cases on the S_{11}. This seems to confirm that, for this DUT, charge trapping has a significant impact on the output- and trans- characteristics of the device even at high frequency, while the impact on the gate charge function seems to be negligible [174].

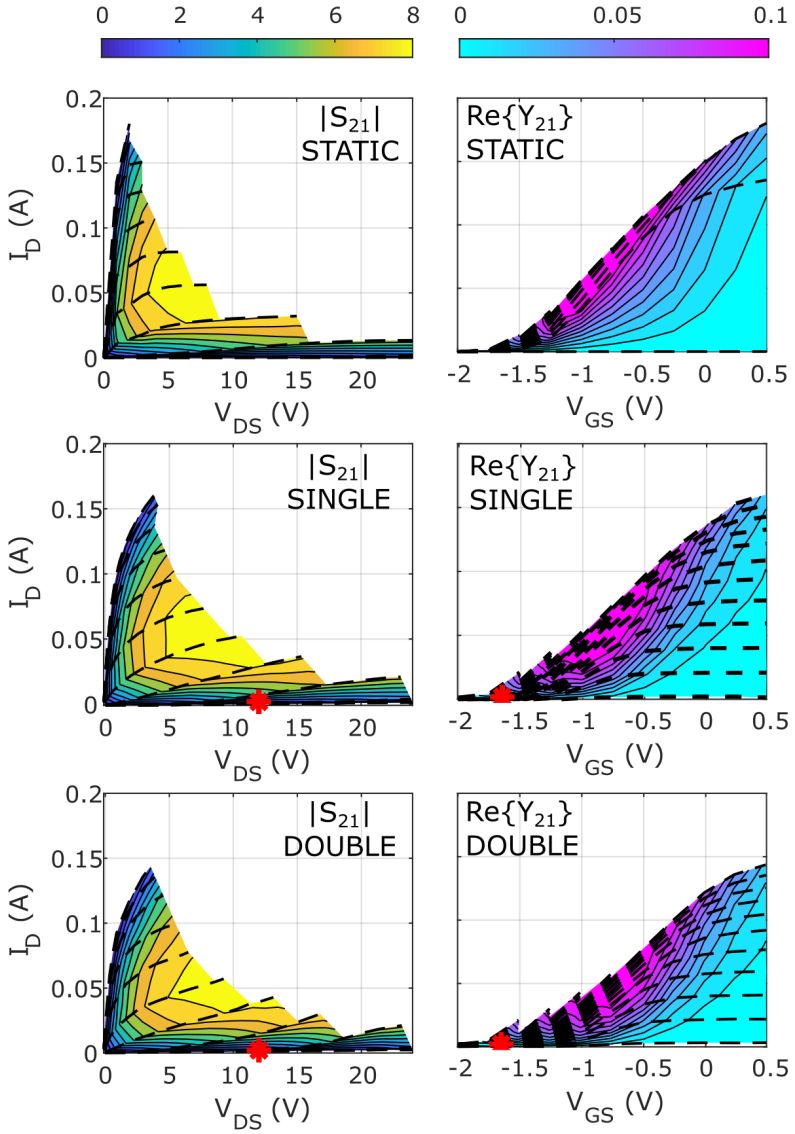


Figure 3.4: Static, single-pulsed and double-pulsed iso-$|S_{21}|$ and iso-$\Re\{Y_{21}\}$ loci at 5 GHz superimposed on the output characteristic and transcharacteristic ($T_C = 40°$ C). Quiescent point $V_{DS,Q} = 12$ V, $I_{DS,Q}$=5 mA in red.

The difference between the single and the double pulsed S-parameters strongly depends on the considered IV point. In point A, the double-pulsed S_{21} is $\sim$ 16 dB, while the single-pulsed one is $\sim$ 20 dB, whereas in point B, the two values have less than 0.5 dB difference.

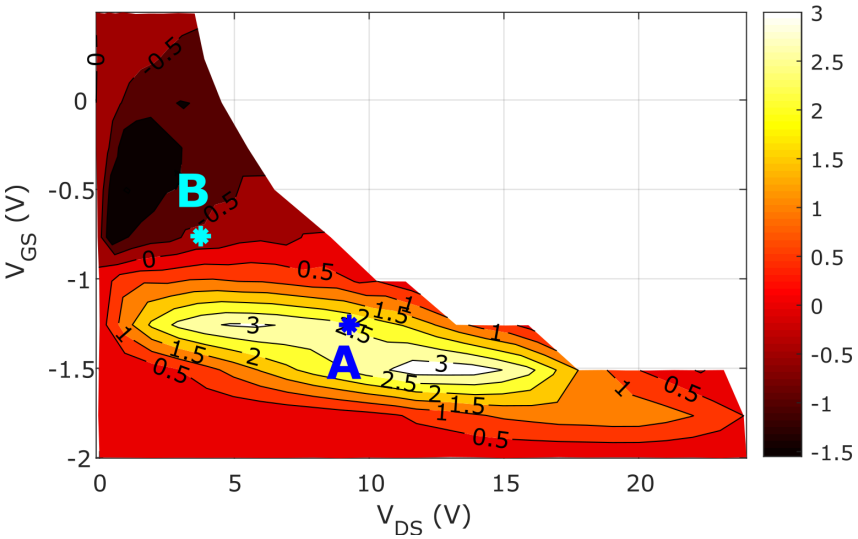


Figure 3.5: Loci in the (V_{GS}, V_{DS}) plane of constant $|S_{21}|$ difference $|S_{21}|_{SP} - |S_{21}|_{DP}$ (in linear units) between the double-pulsed and the single-pulsed cases at 5 GHz (same conditions as in Fig. 3.4). Two specific points in the voltage plane (A and B) are highlighted.

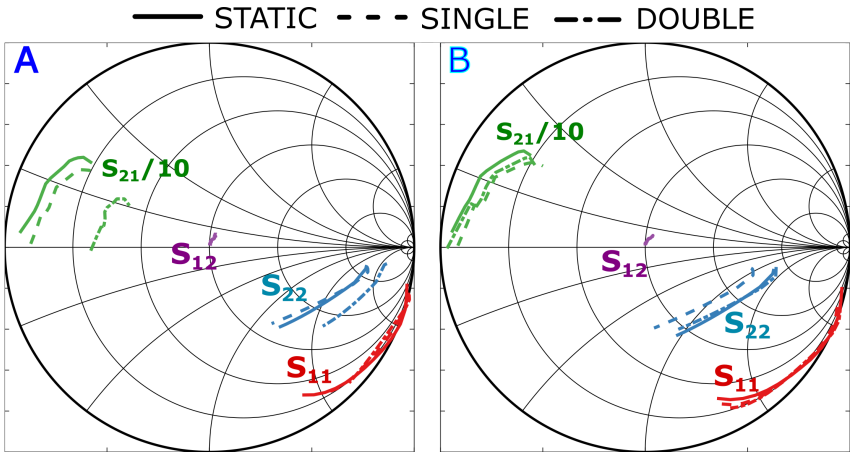


Figure 3.6: Static, single-pulsed and double-pulsed S-parameters for voltage points A ($V_{GS,Q} = -1.25$ V; $V_{DS,Q} = 9.25$ V) and B ($V_{GS,Q} = -0.76$; V $V_{DS,Q} = 3.25$ V) in the 1-7.5 GHz frequency range.

3.3 Microwave characterization of trapping effects in 100-nm GaN-on-Si HEMT technology

3.3.1 Characterization technique

In this Section we examine two different techniques for assessing the impact of trapping dynamics in high-frequency large signal operating conditions: two-tone and RF pulsed

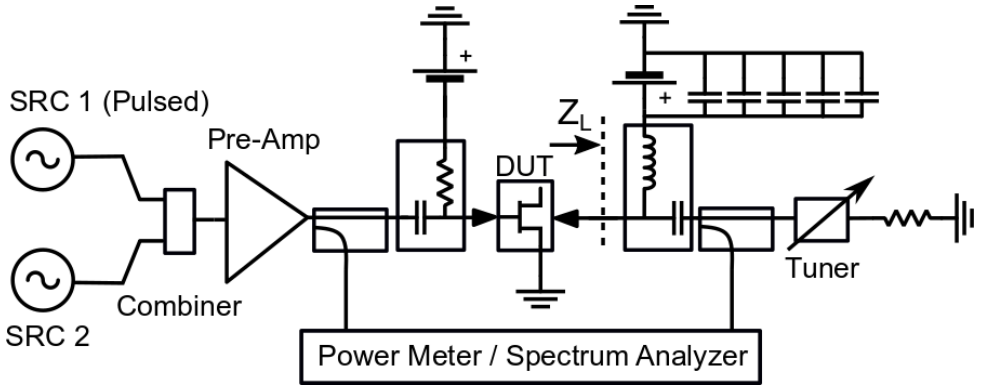


Figure 3.7: Block diagram of the measurement setup.

characterizations. These stimuli, when suitable impedance terminations are applied to the device-under-test in order to enforce a realistic load-line, are able to reproduce, to some extent, the time-varying envelopes observed in application-like scenarios.
Two-tone characterization are standard in high-frequency behavioral modeling [103]. As the frequency spacing between the tones is changed, the amplitude modulated envelope excites different dynamic phenomena, depending on the time-constants in play. The effects are displayed as specific signatures on the intermodulation distortion tones when the frequency spacing is of the same order of magnitude as the inverse of the time-constant of interest.
Pulsed excitations, instead, more closely emulate the operating conditions of radar transmitters or high PAPR signals. Similarly to low-frequency I-DTLS techniques, a short high-power RF pulse is applied at the input of the device [59]. This stimulus rapidly fills the trap levels at a value that depends on the pulse amplitude. The input power is then reduced to a lower, albeit non-zero, value. This allows to observe the slow detrapping process as a long transient in the output power, which progressively recovers to its equilibrium value. The application of a zero-input condition would impede the proposed characterization, as the output power of the device is used in this case as a probe of the internal instantaneous trapping level.

3.3.2 Measurement setup

The measurement setup (Fig. 3.7) is designed to operate at a frequency of 18 GHz. The DUT is injected by two phase-locked RF sources combined with an isolated Wilkinson combiner and then pre-amplified. As one of the two sources has pulsed capabilities, this combined configuration enables both two-level pulse and two-tone sweeps used for the characterizations.
The signals at the input and output ports are sensed by directional couplers, and measured by either a diode-based power sensor for the pulsed case or a spectrum analyzer for the

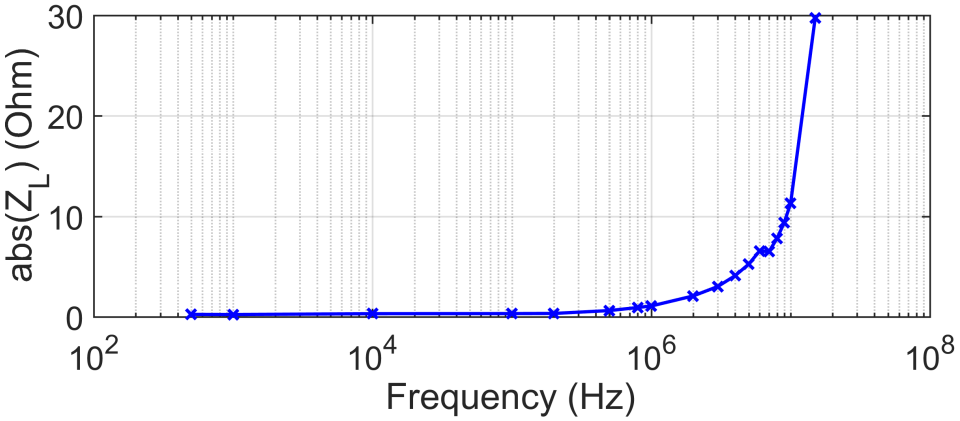


Figure 3.8: Measured load impedance absolute value abs(Z_L) shown by the measurement setup in the low frequency range (f <15 MHz).

two-tone case.

Since trapping mechanisms show nonlinear dynamic dependencies on the applied voltages, maximum care should be paid to the load termination (Z_L) shown by the measurement setup. At the fundamental frequency, a manual tuner has been used for setting the impedance for maximum RF output power, and it was also verified to be flat over frequency in the modulation range. The use of the tuner subjects the device to realistic loadline operating conditions, allowing for large voltage swings to be observed. The setting of an unrealistic termination such as the standard 50-ohm one, might lead to a greatly reduced voltage excursion at the unmatched DUT output, with a consequent underestimation of trapping effects, which have been shown (Sec. 2.3) to be strongly voltage-dependent.

In order to exclude bias-dependent effects [178], the LF termination has to be properly imposed by the setup as well. Indeed, the baseband termination interacts with the dispersive effects modifying the operating regime, and the voltage peak reached by the RF loadline will set the amount of trapped charges [140]. Since the LF dispersion cuts-off below a few MHz, proper termination in the significant modulation range was ensured at baseband by choosing a drain bias tee with relatively high cut-off frequency at the bias port (100 MHz), yet also adding five shunt capacitors from 10 μF to 1 nF to flatten the output impedance of the power supply, obtaining Z_L in the LF range as reported in Fig. 3.8.

The technology under evaluation is the GaN-on-Si process with a 100nm gate length introduced in Sec. 2.3 and in the previous Section. The DUT measured in this work is a 4x50 μm device biased in deep class AB at $V_{DS,Q}$ = 12 V and $I_{DS,Q}$=7.5 mA. At the nominal bias point and matched for maximum RF output power (P_{OUT}), the device shows $\sim$ 27.5 dBm of saturated power at 18 GHz and 9 dB of transducer gain.

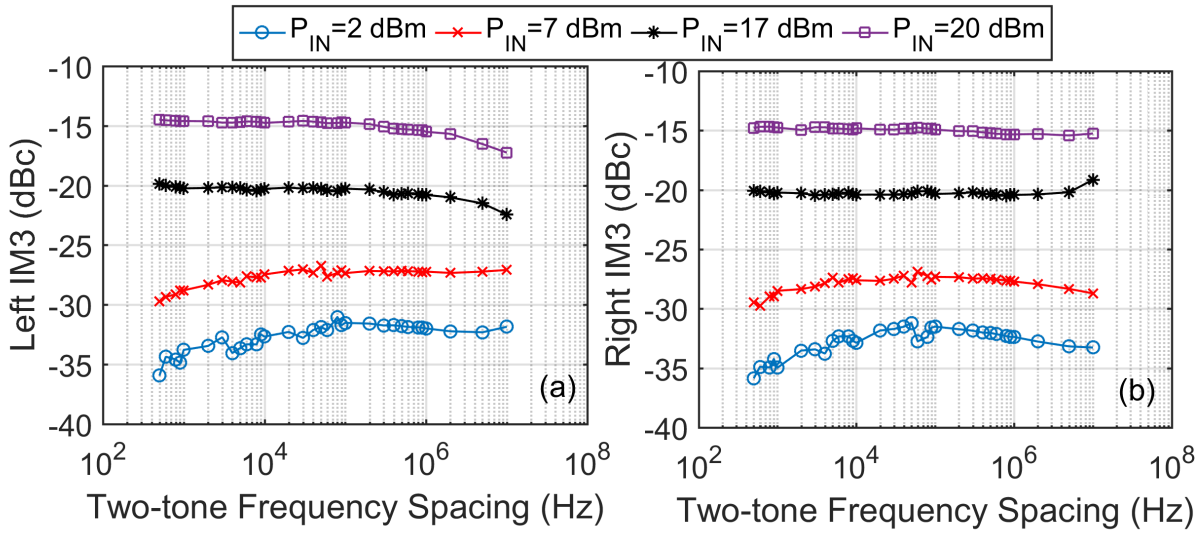


Figure 3.9: Relative left IM3 (a) and right IM3 (b) for RF available input powers 2, 7, 17 and 20 dBm.

3.3.3 Third-order intermodulation distortion measurements

IM3 characterization is widely used to assess both the device nonlinearity and memory effects, and the likelihood of a successful linearization (e.g., by digital predistortion). The presence of a frequency dependency, and of asymmetry between the left and right IM3, gives information on the memory effects, and on the activation at different RF powers. Despite the popularity of the technique, it should be noted that a two-tone signal features PAPR $\sim$ 6 dB, which may be too low with respect to some of the signals used in modern telecom standards and therefore poorly reproduce their statistical properties.
Considering the time-constants of traps in this technology, IM3 must be measured with closely-spaced tones, with spacing frequencies (Δf) well below 1 MHz [151]. These types of measurements are typically challenging to make using network or spectrum analyzers. Indeed, the internal IF BW must be set significantly narrower than the tone spacing of interest, in order to avoid interference between the measured tones images at IF. These narrow BWs are not possible using the fixed analog IF filter banks and make the use of DSP filtering, which greatly increases measurement time. Long measurements are proportionately more affected by LO phase noise (whose spectral density typically increases lower frequencies) and drifts. Similarly, generation of two tone signals at narrow spacings is proportionally more affected by the phase noise of the generating source(s) and in some cases can interfere with the internal output power leveling loops.
Figure 3.9 reports the left and right IM3 measurements (in dBc) at different available input powers of 2, 7 and 17 dBm. The IM3 tones are measured by sweeping Δf from 500 Hz up to 10 MHz around 18 GHz by means of a spectrum analyzer with resolution BW

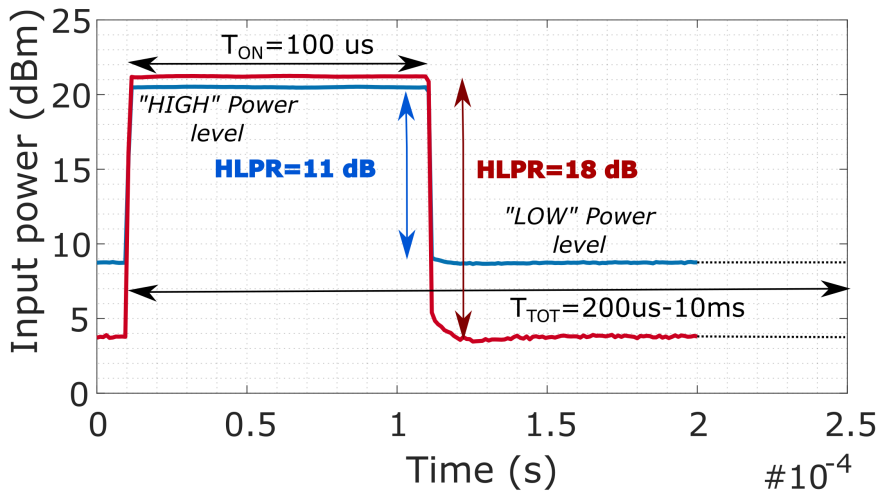


Figure 3.10: Pulsed-RF acquisition at different HLPR levels, obtained by combining two RF CW sources.

$< \Delta f/100$. Both IM3s show very little dependence on the tone spacing, with an overall variation that is more noticeable at lower input powers and for wider spacings (>1 MHz) where the effect of the bias network is probably becoming more noticeable. Similarly, asymmetry between the two IM tones is is not particularly pronounced and seems to diminish with increasing output power. While the technology has been shown to display significant trapping dynamics in the frequency range under examination (Sec. 2.3), the effect on this type of two-tone measurements is hardly noticeable. The slight memory effect seen on the measurements seems to increase at lower power levels. These results contrast with the ones observed for two-tone characterizations of other GaN technologies known to be subject to trapping phenomena, where extremely marked effects can be observed using this technique [103].

3.3.4 Two-level pulsed-RF measurements

Alternatively, the nonlinear dynamic effects can be evaluated by measuring the transient behavior of the traveling waves at the DUT reference plane with a pulsed-RF excitation, testing different powers, duty-cycles (d), or periods. Given the large observation times required by trap recovery, a real-time power meter (Sec. 1.3) is better suited for the transient characterization. Differently from IM3, pulsed-RF with short duty cycles can be considered isothermal, as no substantial self-heating takes place. Here, two combined sources for synthesizing two RF power levels P_{high} and P_{low}, with a high-to-low power ratio defined as $\text{HLPR} \stackrel{\text{def}}{=} P_{high}^{dBm} - P_{low}^{dBm}$, allow for arbitrary PAPRs (Fig. 3.10).
The two-level pulsed-RF characterization has been performed by adjusting the P_{high} input

Table 3.1: Peak-to-average power ratio (PAPR) and estimated time constant (τ) for two-level pulsed-RF excitations

T_{tot} (ms)	0.2	0.5	1	2	3	4	5	10
HLPR=11 dB								
PAPR (dB)	2.7	5.8	7.6	9.0	9.6	9.9	10.1	10.5
τ (ms)	n.a.	n.a.	0.35	0.69	0.83	1.1	1.3	1.5
HLPR=18 dB								
PAPR (dB)	2.9	6.7	9.4	11.9	13.2	13.9	14.5	15.9
τ (ms)	n.a.	n.a.	0.36	0.71	1.1	1.3	1.6	2.7

level to ensure maximum pulsed-RF $P_{OUT} \simeq 27.5$ dBm in all tested regimes, while input level P_{low} is chosen to synthesize either HLPR=11 dB or HLPR=18 dB, obtaining the results in Fig. 3.11. The pulse width (T_{on}) was kept constant at 100 μs, while the period T_{tot} was swept from 200 μs ($d = 50\%$) to 10 ms ($d = 1\%$), resulting in different PAPR values computed as PAPR (dB) $=10\log_{10}\left[\frac{P_{high}}{P_{high}d+P_{low}(1-d)}\right]$ (Table 3.1). After the high-power pulse at P_{high} is switched off, there is a sharp gain collapse during the initial phase with P_{low}. A limited fast ($\sim \mu$s) transient is followed by a longer ($\sim$ ms) recovery, which is a well-known trap signature [170]. The gain collapse, defined as the ratio between the gain after the fast transient and the static CW gain, is up to 3.5 dB for HLPR=11 dB and almost 6.5 dB for HLPR=18 dB. In both cases, the gain collapse increases for larger values of the duty cycle. The CW gain is larger than the one in pulsed operation for all test cases, indicating that the recovery transient is not completely extinguished, even for the longest tested period (10 ms). For HLPR=11 dB the pulsed gain under P_{high} is inversely correlated with respect to the duty-cycle, while a smaller corresponding variation is observed for HLPR=18 dB. For estimating a characteristic time constant (τ) for each regime, a single exponential function is fitted to the experimental data for RF P_{OUT} (Table 3.1). The recovery time constants for HLPR=18 dB are generally larger than the ones in the case of HLPR=11 dB, with the difference becoming higher for longer T_{tot}. The long-term transient is practically absent during the short pulse OFF-time for T_{tot}=0.2 and 0.5 ms, given that the pulse OFF-time is too short for any trap recovery.

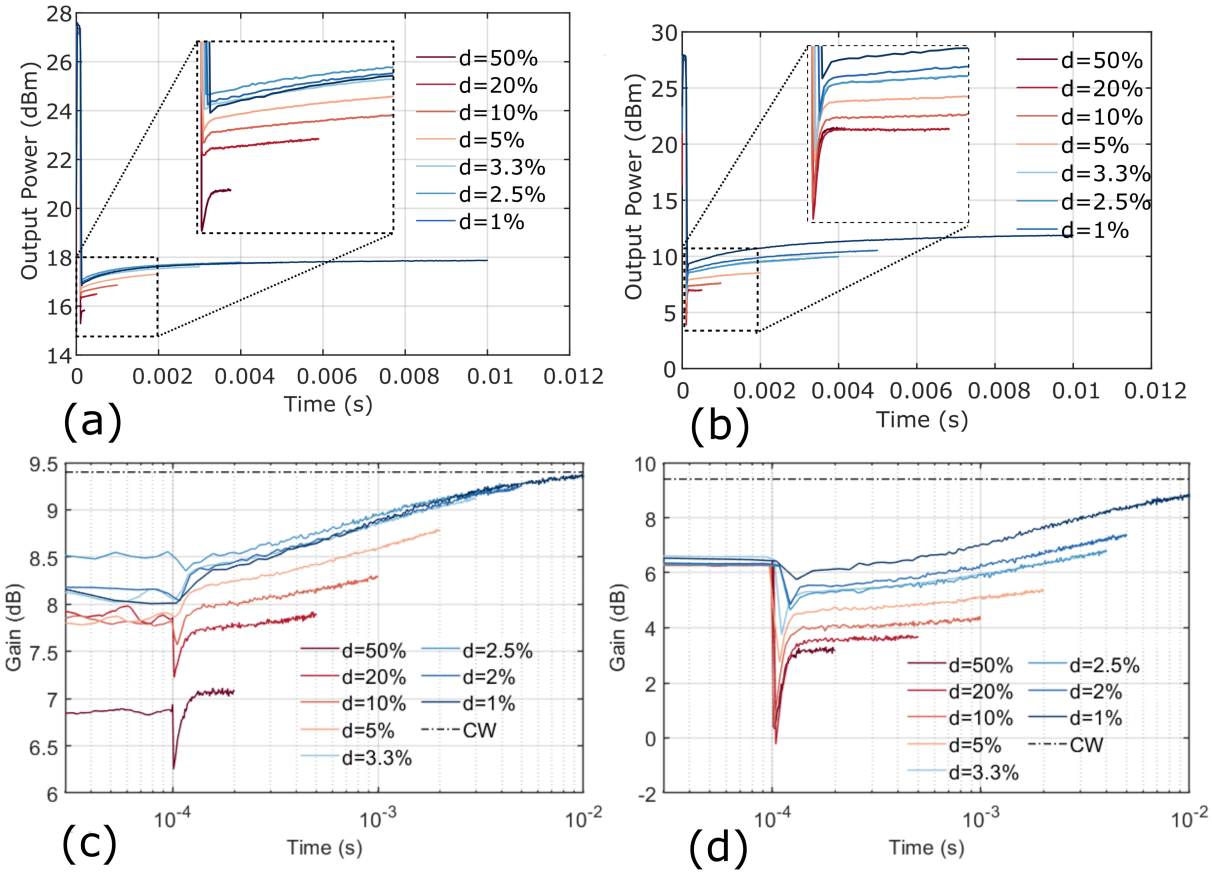


Figure 3.11: Two-level pulsed-RF output power at different duty cycles (d) with T_{on}= 100 μs and T_{tot}= 0.2 ÷ 10 ms. (a) HLPR=11 dB, P_{high}= 19.5 dBm and P_{low}=8.5 dBm. (b) HLPR=18 dB, with P_{high}=21.2 dBm and P_{low}=3.2 dBm. Two-level pulsed-RF gain in the same conditions. (c) HLPR=11 dB, P_{high}= 19.5 dBm and P_{low}=8.5 dBm. (d) HLPR=18 dB, with P_{high}=21.2 dBm and P_{low}=3.2 dBm. Dashed black line represents the measured static CW gain of the device under P_{low} input power.

3.4 Broadband EVM characterization of a GaN power amplifier using a VNA

3.4.1 Characterization technique

EVM is a key-metric in characterizing the broadband linearity performance of RF and microwave PAs in communication transmitters. As linearity is critically dependent on the PA large-signal operating point (LSOP), it is of great interest to evaluate how different input signal characteristics might influence the PA performance in this respect.
For example, the dependence on the input RMS power has been studied in detail [116] in

order to generate EVM tradeoff curves, which can give an indication of wideband distortion levels that can be expected across various backoff levels. On the other hand, while it has been shown that the LSOP depends on the periodicity of modulated excitations in presence of significant long-memory effects [141], [179] in GaN PAs, the effective impact on the overall linearity has rarely been reported [180].
In this light, the present Section investigates the possibility of detecting the direct effect of LF long-memory phenomena on a system-level RF metric such as EVM. This, in principle, allows to quantify what is the expected deterioration in transmitter performance due to charge trapping in GaN devices that are commonly used to build the PA. Moreover, differently from the characterizations proposed in the previous sections, the excitations used for EVM measurement can be specifically tailored to match pre-defined standards of interest, providing a very close approximation of the real-world operating conditions.
This evaluation can be done using periodic multitone signals that approximate the OFDM stochastic process of interests for the 5G standards [3]. Provided the total number of tones is in any case sufficiently high to provide good approximation properties [113], signals sharing the same asymptotic distribution and bandlimited power spectral density can be designed using different tone spacings [179]. Similarly to the two-tone case in Sec. 3.3, multi-tone modulated signals with different tone spacings excite dynamical effects in the DUT across different time scales, while having a fixed and well defined amplitude distribution for all cases. As fast-capture occurs during RF power peaks and slow-release is triggered at lower power levels, the complex dynamics under realistic modulated conditions can be measured in terms of the impact on the EVM performance of the PA.

3.4.2 Measurement Setup

The setup used in this work (Fig. 3.12) is composed by a broadband Vector Signal Generator (VSG), Keysight N5182B MXG, a pre-amp (Mini-Circuits ZVE-3W-183+) for properly amplifying the a_1 input signal, and a VNA, Keysight N5242A PNA-X, which is calibrated by a classic short-open-load-thru relative calibration with connectorized standards. Also, a power calibration referenced to a power meter is employed to correctly measure the tone amplitudes. The maximum measured modulation bandwidth, centered at 5.5 GHz, is 200 MHz, corresponding to the maximum BW of the VSG. Nevertheless, there is no specific limitation on BW on the receiver side, up to the 10 MHz - 26.5 GHz range of the VNA test-set. The DUT is a monolithic two-stage broadband PA in 250-nm GaN-on-SiC technology biased at $V_{DS} = 20$ V and $V_{GS} = -3.20$ V. The technology is known to display significant trapping phenomena due to doping the in buffer layer [28] and is therefore a promising candidate for the reported characterization.

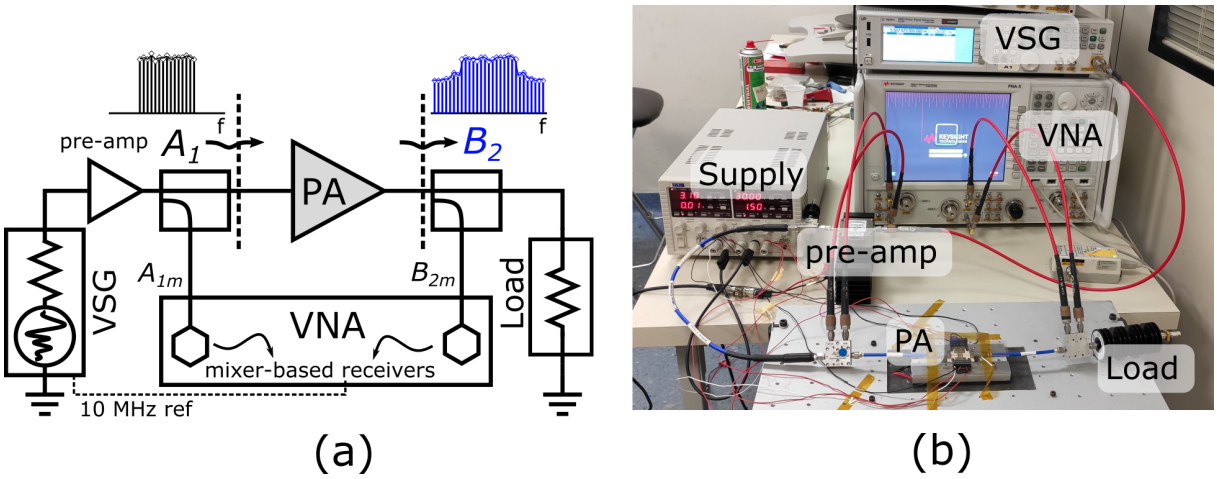


Figure 3.12: (a) Block diagram of the VNA-based measurement setup. (b) Photo of the setup.

3.4.3 Best Linear Approximation and VNA-based Broadband EVM Estimation

According to the BLA theory introduced in Sec. 1.4, for any input signal A_1 in the gaussian class of signals with a standard-like pdf and PSD, the output spectrum B_2 can be written at each frequency as [114]:

$$B_2(f) = G_{BLA}(f)A_1(f) + D(f) + N(f). \tag{3.1}$$

In 3.1 G_{BLA} is the BLA gain of the amplifier, $D(f)$, with variance $\sigma_D^2(f)$, quantifies the input-uncorrelated nonlinear stochastic distortions, while $N(f)$, with variance $\sigma_N^2(f)$, represents additive measurement noise on the output. After the measurement of these quantities, EVM can then be estimated as the ratio between purely nonlinear distortion power (i.e., the output signal power minus the linearly input-correlated power) and input-correlated signal power across the excitation bandwidth:

$$EVM = \sqrt{\frac{\int_{BW} \sigma_D^2(f)df}{\int_{BW} |G_{BLA}(f)|^2 S_{A_1A_1}(f)df}} \tag{3.2}$$

where $S_{A_1A_1}(f)$ it the measured input PSD.
Fig. 3.13a shows an example of the BLA characterization method applied to the GaN PA under test, showing the distortion contribution to the output signal. The measurements were performed using a 20-MHz wide random phase multitone with 3 kHz spacing and and input power of 11.2 dBm, using $P = 2$ periods and $M = 25$ phase realizations for estimating the stochastic expectations. The measured G_{BLA} together with its nonlinear stochastic and noise variance components is shown in Fig. 3.13b.

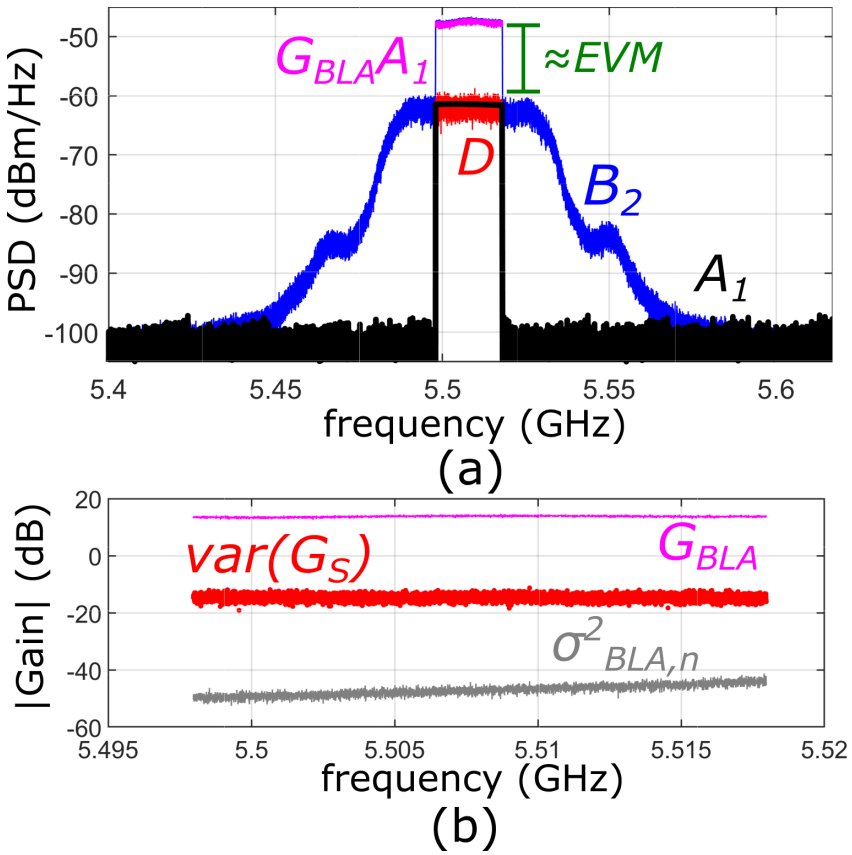


Figure 3.13: a) Estimated input (black) and output (blue) power spectra for an input RMS power of 11.2 dBm in case of a 3 kHz-spaced 20-MHz-wide random phase multitone excitation. Two measurement periods and 25 phase realizations are used. The correlated output (magenta) and uncorrelated distortion (red) spectra used in the EVM computation are outlined. b) Estimated BLA gain for the amplifier (magenta), together with the noise variance (gray) and the variance of the stochastic nonlinear contributions (red).

3.4.4 GaN PA EVM Characterization

Figure 3.14 shows the EVM and RMS ouput power as a function of the input RMS power and tone spacing for the amplifier under exam. Reference EVM levels for different carrier-modulation formats as per 5G [3] specifications are added to the plot, in order to evaluate which types of constellations are compatible with a given distortion level. Two FR1 5G-compliant signal BWs of 20 and 100 MHz have been tested, with the wider bandwidth remarkably displaying a lower EVM and higher linearity for the same input power. Nevertheless, in both cases, EVM and output power seem to be negligibly affected by the a variation of the tone spacing from 3 KHz up to 150 kHz.
Figure 3.15 reports the measured BLA as a function of input power and tone spacings for

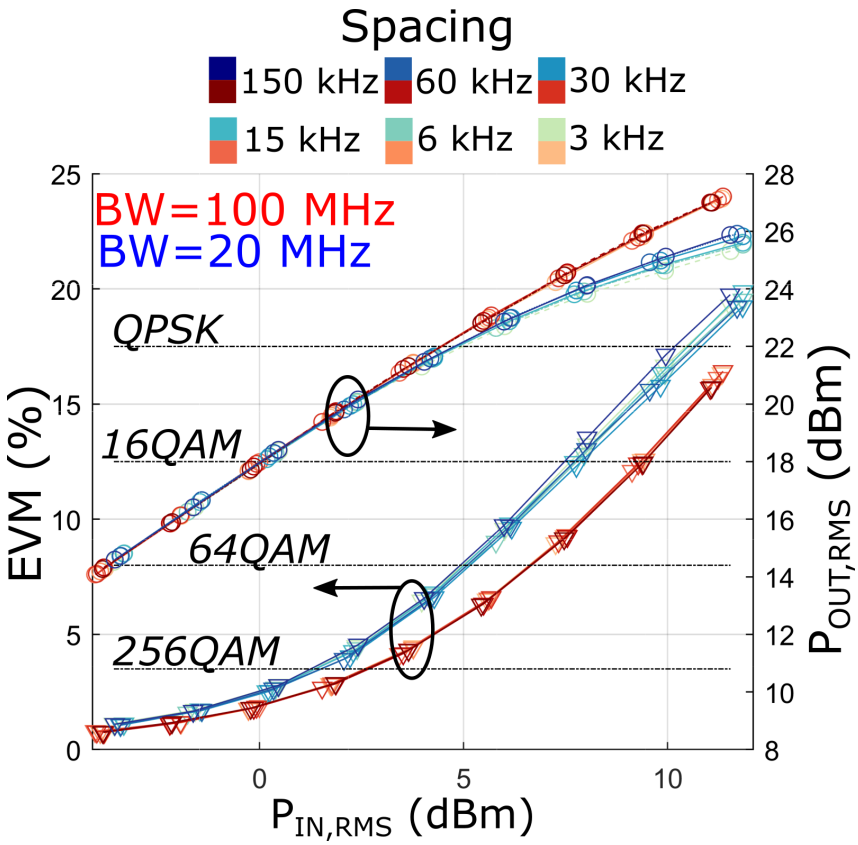


Figure 3.14: EVM and output RMS power for the amplifier under test as a function of the input RMS power. Two different random phase multitone bandwidths (20 and 100 MHz) and six different tone spacings (3 kHz to 150 kHz) are reported. Reference EVM levels for different modulation formats as per 5G specifications are superimposed.

the two examined bandwidths. The estimated $G_{BLA}(f)$ profile across frequency displays a marked shape variation across the input power levels and a significant difference between the two excitation bandwidths. Indeed, the BLA is a pdf-and-PSD-dependent combination [114] of the small signal linear response (seen at low input power) with higher-order Volterra kernels, which are clearly relevant for the LSOPs under consideration.
For both cases, the BLA gain of the PA generally decreases with power, similarly to what is observed in static gain-compression characteristics [99]. For the 20 MHz case, this decrease is sharper at the band edges, giving the overall frequency profile a concave shape at higher powers. Conversely, for the 100 MHz bandwidth case in the same operating conditions, the profile becomes convex, with higher reduction in gain around the 5.5 GHz carrier. This points to a rich nonlinear dynamic behaviour for the PA, which cannot be described or efficiently linearized using widespread memory-polynomial or Wiener-Hammerstein structures [181]. Indeed, for these simplified models, the BLA is known [63]

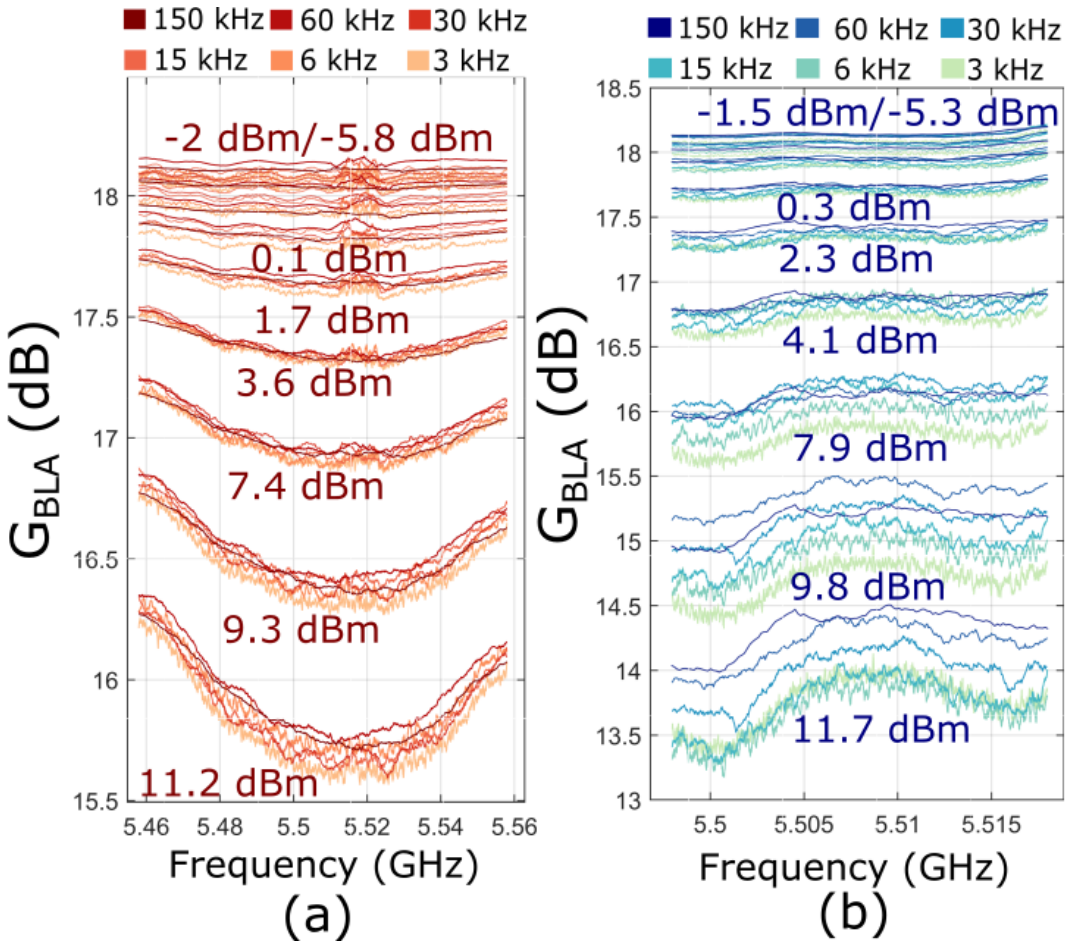


Figure 3.15: Magnitude of the estimated best-linear approximation for the amplifier under test as a function of the input RMS power and tone spacing (3 kHz to 150kHz). Signal bandwidths of a) 100 MHz and b) 20 MHz are reported.

to display a specific signature, maintaining the exact same shape even at different input power levels. The variations of the BLA for different tone spacings is negligible in the 100 MHz case, while some minor but measurable variations of ~ 0.5 dB can be seen for the 20 MHz BW.

Dc current of GaN HEMTS is known to be particularly sensitive to trapping state variations [141] and is therefore a prime candidate to gain valuable information on the LSOP of the DUT. In this characterization, it can be used to assess if the differences in the BLA behavior in the 20 MHz case are indeed due to captured charge variations at different tone spacings. The dc current, averaged across all the periods and realizations of the input signal, is reported for each input power and tone spacing in Fig. 3.14. The two signal bandwidths display remarkable differences in behaviour. The 20 MHz BW displays

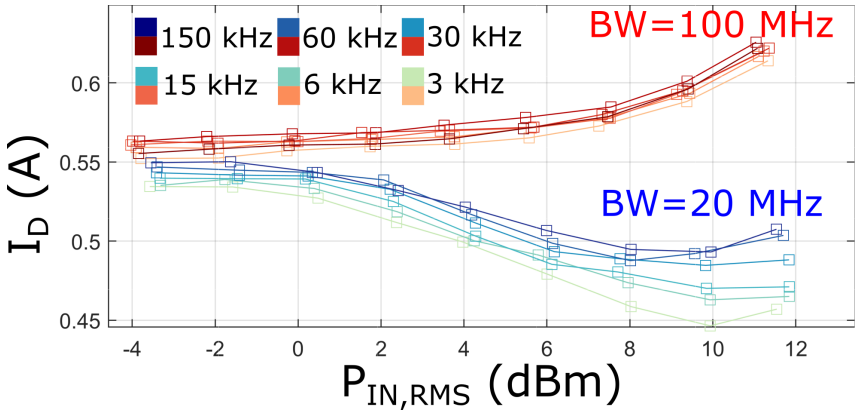


Figure 3.16: Dc drain current for the amplifier under test as a function of the input RMS power. Two different random phase multitone bandwidths (20 and 100 MHz) and six different tone spacings (3 kHz to 150kHz) are reported.

a relevant current collapse with increasing power, which is a well known trap-induced signature [141]. Moreover, the collapse significantly depends on the tone spacing, pointing to different dynamics of the effects for the observed LSOPs . On the other hand, the effect is not observed for the wider excitation BW.
This behavior matches the one observed for the previous measurements, giving an indication that trapping might be indeed a possible cause of the observed difference between the BLA gains at different tone spacings in the 20 MHz case. At the same time, they seem to confirm that no significant trapping and dynamic differences can be observed in the 100 MHz case.
The reported results cannot rule out that more pronounced dependencies of the EVM/BLA on the tone spacings could arise if narrower spacings (i.e., longer memory effects) were considered for the proposed characterization. Indeed, detrapping time constants have been shown to exceed several seconds in some cases [126]. However, this possibility has to take into account that in typical receivers, a new equalization/BLA is computed for each 10 ms frame [3]. This means that if the DUT time constants are longer than the frame length, the trapping dynamic behaviour is seen at the receiver side as a long-term drift of the DUT response that can be compensated with an update of the equalizer coefficients. So, extremely slow (10 ms-s range) memory transients should not have a significant impact on the overall EVM in a realistic transmission-reception chain.

3.5 Conclusions

In this chapter, several different characterization techniques for the assessment of the impact of trapping effects on the RF characteristics of GaN HEMTs have been reported.

The implementation of a novel high frequency measurement bench has enabled the characterization of double-pulsed S-parameters of GaN HEMT. On top of the measurement of isodynamic RF differential parameters, it allows the assessment of the trap-induced degradation of S-parameters, showing relevant differences between the dc, single- and double-pulsed cases. The importance of the observed effects strongly depends on the considered operating point, and is observed to affect mostly the S_{21} and S_{22} parameters. This type of evaluation.

Large signal RF trapping characterizations of a novel 100-nm GaN-on-Si technology have been reported. While a two-tone measurement at different spacings does not show any clear signature that trapping might be relevant, the DUT has revealed a clear transient behavior due to traps when matched for maximum RF output power and excited by high-PAPR pulsed-RF signals. Under the practical operating conditions tested in this work, the recovery time increases monotonically with the PAPR, with regime-dependent time constants ranging from a few hundred μs to a few ms, which are comparable to what observed in low frequency characterizations of the same technology.

Finally, a novel characterization method to evaluate the impact of long-memory dynamic phenomena, and particularly trapping, on the wideband linearity of GaN PAs has been proposed. The method involves the extraction of the best-linear approximation of the PA using standard-emulating multitone signals at varying tone spacings in order to probe the memory behavior across different time scales. For the examined PA, EVM shows little dependence on the actual tone spacing of the multisine used in its evaluation, despite the fact that the technology is known to exibit significant charge trapping effects. At the same time the LF dc current and the extracted BLA gain indicate that that trapping might be indeed at play in determining the device LSOP. In any case, measurements show a behaviour that strongly depends on the modulation bandwidth and input rms power, pointing to complex nonlinear dynamic behaviour to be accounted for during system level modeling and linearization.

These research results highlight the requirement for RF large signal characterizations techniques in the assessment of trapping-induced performance degradation of microwave electron devices. Even when low-frequency characterizations indicate the presence of trapping, the observed effects during operation at RF frequencies might show conflicting signatures that are strongly dependent on the excitation and the overall large signal operating point. This constitutes a challenge for many well-known HEMT models, which, on top of the charge capture and release dynamics have to describe this complex interaction between LF and RF phenomena. In order to improve predictive performance, it is therefore of paramount importance to extract and validate these models in RF operating conditions that closely mimic the ones encountered in the final application.

Chapter 4

Wideband Load Pull Characterization Techniques

4.1 Introduction

As introduced in Sec. 1.3.7, wideband active load pull has been proposed as a promising technique to characterize microwave transistors and to optimize PA performance using excitations that mimic typical communication standards [75], [182]–[184]. The method allows to set, for a given device-under-test excited by a broadband modulated signal, a user-prescribed load reflection coefficient profile across a wide BW at the fundamental frequency, and possibly at its harmonics. This type of capabilities are fundamental in characterization of GaN HEMTs, as the severity of trapping phenomena has been shown to strongly depend on the large signal operating point, and therefore on the broadband loading conditions, of the DUT.
Several commercial solutions for WALP are available [75][184], which make use of time-domain waveform demodulation capabilities exploiting broadband acquisition hardware [75], [182]–[184]. Section 4.2 is devoted to investigate the possibility of enabling WALP under modulated excitations by leveraging on standard VNA measurements, first published in [185], [186]. This approach allows for WALP by using HW commonly found in microwave labs, and it completely avoids the need of a traceable phase reference standard or absolute phase calibrations. Moreover, it removes the need of single-shot IQ demodulation of large instantaneous BWs, enabling WALP on the much wider excitation BWs required by modern communication standards.
These capabilities enable several VNA-based nonlinear measurements [113], [173], [180], [187] directly at transistor level. Indeed, the proposed VNA-WALP allows to emulate the broadband output termination conditions seen in the final application, allowing to characterize device trapping and memory behavior under realistic modulated excitations. This permits to efficiently characterize several key performance metrics inherent to the

technology at an early stage of design, before the physical realization of any actual PA circuit.
The aim of the proposed VNA-WALP is to set any arbitrary load reflection profile in frequency domain, without depicting any particular PA design strategy, or the identification of a single specific broadband optimal termination. It is important to notice that, in the wideband case, it is typically unfeasible to test a large number of loads at each frequency in order to identify a suitable load impedance that optimizes device performance. Indeed, if N frequencies are considered in a load pull sweep grid with L different loads at each frequency, load pull based on a full-factorial design of experiment would require presenting all the L^N resulting profiles to the DUT. While traditional narrowband load pull counts only a few frequency points (a fundamental frequency point and possibly a very limited number of harmonics), practically allowing for an extensive exploration of the design space and to obtain the well-known contour plots, WALP across broadband modulation usually counts thousands of frequency points, which do not easily permit such an extensive design space exploration. The exploration can be limited to a subset of this search space, for example by considering the same load point for all the frequencies in the bandwidth of interest. Nevertheless, the setup has the capabilities to explore other configurations, such as the emulation of different electrical delays, that are unavailable in traditional narrowband load pull.
Instead, in Sec. 4.2, the theoretical and practical possibilities of performing EVM characterizations in a non-50 Ω environment are explored. As outlined in Sec. 1.4, EVM characterizations have been enabled in VNA-based measurement benches [113], with capabilities that are compatible with those required by the proposed VNA-WALP setup. In this respect, the combination of the two techniques could allow the measurement of the EVM at different load conditions. This has indeed been shown [75] to be a valuable tool for characterizing the broadband distortion performance of PAs when an output mismatch is encountered, such in the case of an antenna or a time-varying impedance due to coupling between different elements. Another interesting application is in the assessment of the intrinsic wideband linearity of on-wafer microwave transistors, where the distortion on a actively-synthesized load can be used to assess final RF PA performance at an early stage and without any physical implementation requirement.
The main challenge in this measurement lies in the presence of several possible signals of interest, while EVM is typically referred to the single-input and single-output (SISO) case [116]. Indeed, in a load pull scenario, both incident and reflected waves at different reference planes are all simultaneously non-zero, raising some doubts on which are the correct signals (e.g., a_2 or b_2) to consider in order to compute a meaningful distortion metric. The implications on this type of EVM characterization are theoretically analyzed by studying the general system obtained by the cascade of a nonlinear-dynamic block with a linear two-port network. Then, the examined situation is experimentally reproduced by using an RF power amplifier, with a load consisting of a passive slotted-line tuner.
Conclusions are drawn in Sec. 4.4.

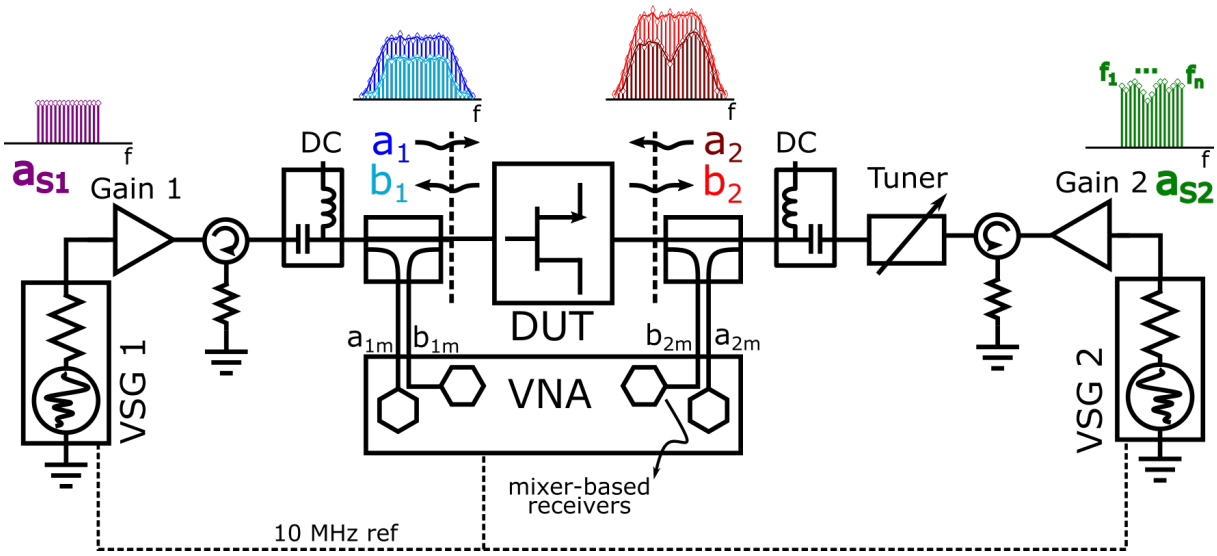


Figure 4.1: Block diagram of the WALP setup.

4.2 Wideband active load pull using a vector network analyzer

4.2.1 WALP mathematical formulation

Let us consider the case in which the DUT is excited at the input by a user-defined periodic band-pass signal $a_{S1}(t)$ around a single RF carrier. This general scenario represents many relevant application cases in PA stimulus-response characterization measurements. In particular, this work will focus on modulated excitations, such as the ones used for 5G. Indeed, periodic multitone signals can be designed to mimic the statistical and spectral properties of communication standards [64]. Among the many different techniques available in literature [5], [64], [113], [188], flat-amplitude random-phase multitones with a prescribed BW will be used in this work to provide a close approximation of complex-gaussian envelope of 5G-OFDM waveforms.
Due to the nonlinear distortion introduced by the DUT, the traveling voltage waves at the input and output reference planes will be composed of several tones in frequency domain around the fundamental and harmonics. These tones, under the reasonable assumption that the DUT displays PISPO behaviour [112], will fall on a predictable frequency grid with the same spacing as the original excitation signal a_{S1}.
While the baseband as well as harmonic source and load terminations may have a significant impact on the device behavior [75], [183], this work will only consider load pull for a certain BW around a given RF carrier. This is still a relevant test case in many applications, particularly at higher frequency ranges where harmonics are difficult to measure

and/or control both in the characterization and design stages [189]. In all cases, the implementation of a full multiharmonic and baseband WALP setup entails the use of a substantial amount of hardware and software capabilities.
Overall, the frequency domain measurements in the proposed setup will include both the input-excited tones and the spectral regrowth such as third or higher order intermodulation (IM) distortion products. The reflection coefficients seen by the DUT at baseband and harmonic frequencies will not be directly controlled by the setup, and will coincide to those presented by the adopted test-set (e.g., couplers, bias-tees, sources, tuners) at the respective reference planes.
The user selects N tones of interest at the frequencies $f_1 < f_2 < \cdots < f_N$, which must lie on the multitone frequency grid, across which a given target output reflection coefficient $\Gamma_T(f)$ has to be imposed by the setup. In order to synthesize this load profile across frequency, the strategy consists of actively injecting power at the frequencies $f_1 < f_2 < \cdots < f_N$ on the load plane of the DUT using a second signal source. The source and load signal injection sources, under the hypothesis that only the BW around the fundamental frequency is considered, can be realized using two Vector Signal Generators (VSG) with sufficient BW, tuned in the same frequency range, as shown in Fig. 4.1.
The overall goal of WALP is to find the correct signal a_{S2} to inject on the load side, such that the required $\Gamma_T(f)$ is synthesized. In the classic WALP implementation, the computation is typically performed in an iterative fashion, starting from a 50-Ω environment when the load source is turned off, and using successive time-domain waveform acquisitions to guide convergence to the target [75], [182]–[184], [188]. In particular, at a given iteration, the injection signal $a_{S2}(t)$ is computed as the inverse Fourier transform of $B_2(f) \cdot \Gamma_T(f)$, where $B_2(f)$ represents the frequency spectrum of the measured $b_2(t)$. Using successive signal re-injections, the operating regime of the DUT will converge, thanks to the fixed-point theorem, to the stable condition in which $A_2(f) = B_2(f) \cdot \Gamma_T(f)$.

WALP-on-a-VNA

Differently from the available WALP setups, here we explore the possibility of performing WALP using a VNA, without any time-domain waveform reconstruction capabilities, hence without measuring a phase-coherent $B_2(f)$ and its time domain transform $b_2(t)$. Instead, just calibrated amplitude spectra and same-frequency ratioed measurements are exploited in order to set the required target. The use of this reduced subset of measurement prevents the adoption of standard waveform-based algorithms and requires a novel approach to the problem.
Ultimately, the VNA-WALP functionality should impose that the measured reflection coefficient $\Gamma_2(f)$, i.e., the VNA-calibrated ratio between $A_2(f)$ and $B_2(f)$ at the output

reference plane, equals the target $\Gamma_T(f)$ across the N frequencies of interest:

$$\begin{aligned} \Gamma_2(f_n) - \Gamma_T(f_n) = 0, \qquad n = 1 \dots N; \\ \Gamma_2(f_n) \stackrel{def}{=} \frac{A_2(f_n)}{B_2(f_n)} = G_n \left(A_{S2}\left(f_1\right), \dots A_{S2}\left(f_N\right)\right). \end{aligned} \tag{4.1}$$

The load reflection coefficient at any given frequency $\Gamma_2(f_n)$ is a nonlinear function G_n of the injected complex phasors at all the frequencies $f_1 \dots f_N$, as the IM distortion between the injected tones will cause cross-frequency coupling in the A_2 and B_2 waves. By defining the frequency-domain vector $\overline{A}_{S2} \stackrel{def}{=} [A_{S2}(f_1) \dots A_{S2}(f_N)]^t$, (4.1) can be recast as:

$$E_n(\overline{A}_{S2}) \stackrel{def}{=} G_n(\overline{A}_{s2}) - \Gamma_T(f_n) = 0, \qquad n = 1 \dots N; \tag{4.2}$$

where E_n is the reflection coefficient error at frequency f_n. Equation (4.1) can be turned into a system of N complex non-linear equations in N complex variables:

$$\overline{E}(\overline{A}_{S2}) = \left[E_1(\overline{A}_{S2}) \dots E_N(\overline{A}_{S2})\right]^T = \overline{0}; \tag{4.3}$$

where $\overline{E}(\overline{A}_{S2})$ represents the complex error vector between the actual and target output reflection coefficients for a given injected signal.

Finding the correct load injection signal $a_{S2}(t)$ amounts to determining the unique solution (i.e., the zero or root of $\overline{E}$) to (4.3). The error function, in the general case, will not have an explicit expression, as this type of nonlinear frequency-domain mapping can be extremely complex to compute [97] and features a strong dependence on the LSOP. Instead, its value for a given injection can be obtained just through measurements of the DUT.

Finding a zero of a $N \times N$ nonlinear function is a classical problem in mathematics, with several numerical approaches reported in literature [190]. All known general-purpose algorithms are iterative in nature: in order to find a good candidate solution, an algorithm needs to evaluate the function, whose zeroes are to be found, at several different points. Then, using the values at these points, a new *informed* guess of the solution is computed, with the iterations progressing until a suitable stopping criterion is met. For the VNA-WALP, this amounts to computing the load reflection coefficient error $\overline{E}$ for a sequence of different injected signals $\overline{A}_{S2}$. In this context, it should be noted that the reflection coefficient iteratively synthesized by active injection cannot be generally assumed to be smooth across frequency, so no extrapolation techniques like [191] can be exploited.

After the reference input multitone signal a_{S1} is uploaded on the input VSG at the beginning of the load pull sweep, each evaluation consists of two steps. First, using a frequency-to-time transformation to generate in-phase and quadrature (IQ) complex samples from $\overline{A}_{S2}$, the signal $a_{S2}(t)$ is uploaded on the output VSG. Then, the synthesized $\Gamma_2(f)$ at the output reference plane is measured at each frequency f_n and compared to the target in order to compute the error $\overline{E}$. Thus, the acquisition consists of a classical VNA measurement sweep (see Sec. 4.2.2) across all the frequencies at which load pull is performed.

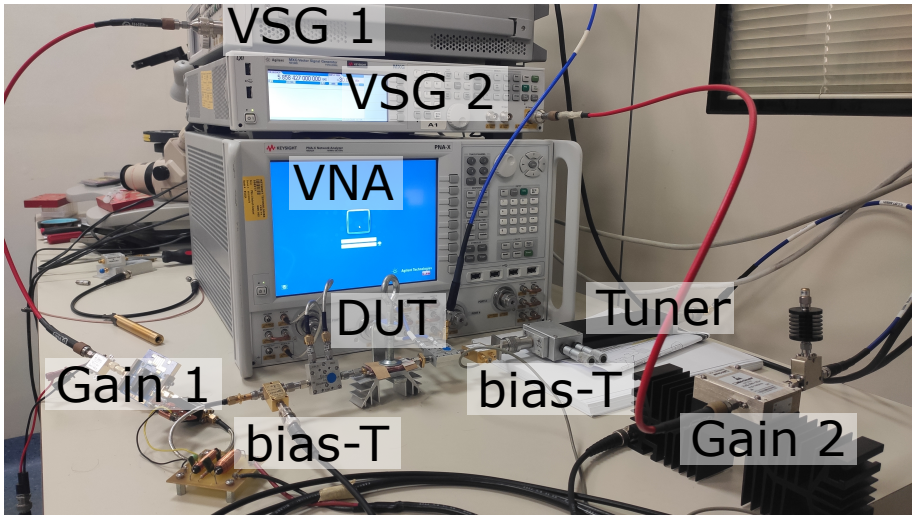


Figure 4.2: A picture of the proposed VNA-WALP measurement setup.

Hence, the total number of evaluations required has a direct impact on the total measurement and convergence times, and using a solving algorithm that can find a suitable solution using the minimum number of function evaluations is of paramount importance (see Sec. 4.2.3).

4.2.2 Measurement setup

The block diagram and picture of the WALP setup used in this work are shown in Figs. 4.1 and 4.2, respectively. The DUT is excited at the input and output ports by two RF VSGs (Keysight MXG-N5182A and MXG-N5182B, respectively). The HW characteristics of the signal generators allow for an input excitation and output load injection real-time BW up to 100 MHz and 160 MHz, respectively, around a frequency carrier up to 6 GHz. The two sources are phase-locked using a 10 MHz reference signal, and the baseband generators are synchronized in time in a master-slave configuration using waveform markers. The output power levels of the VSGs are boosted using two high-gain amplifiers, whose outputs are decoupled from the DUT using circulators. These amplifiers should display highly-linear operation across the whole power range of interest in order to avoid unwanted nonlinear distortion between the digital signals stored in VSG memory and the actual waveforms at the DUT reference planes. While minor nonlinear components can be automatically compensated during the iterative load-setting procedure (Sec. 4.2.3), the use of highly distorting booster amplifiers can indeed jeopardize the convergence of the proposed WALP methods. A passive slotted-line tuner can be added at the output side in order to provide combined passive and active hybrid load pull capabilities.
The acquisition of the four traveling voltage waves is performed by a VNA (Keysight

N5242 10 MHz - 26.5 GHz PNA-X), which is vector (using the short-open-load-thru algorithm) and power (referenced to a traceable power source) calibrated to the DUT connectorized reference planes. As outlined in Sec. 4.2.1, only the same-frequency relative measurements are used in developing the load pull functionality proposed here, with the amplitude calibration used only to set compliance and reference power levels. It is important to highlight that this approach does not make use of any cross-frequency phase reference, as the arbitrary and unknown phase shift of the local oscillator (LO) signal is eliminated when calculating the ratio of synchronous acquisitions. Moreover, standard VNA HW can be used, eliminating the need for the broadband demodulators used in vector signal analyzers [75], [184] to cover the BW of interest.
The acquisitions are performed across a 200-MHz total measurement BW with a minimum tone spacing of 3 kHz. This choice enables power and vector calibrated measurements of up to fifth-order IM distortion spectral components for the 40 MHz-wide excitations that will be used throughout this work to validate the proposed methods. Nevertheless, the setup is capable of working up to the full BW of the instrument front-ends, with arbitrarily narrow tone spacing. These characteristics are typically required in order to perform WALP using multitone signals featuring a good statistical and spectral approximation of the target telecom signal envelope [113], [187], [188]. In the proposed setup, the minimum achievable tone spacing is ultimately limited by the maximum allowed measurement time, available instrument memory, source phase noise, and thermal drifts. All software routines for instrument control and for processing WALP algorithms are implemented in MATLAB®, running on an external LAN-networked computer.

4.2.3 WALP methods' comparison

The performance of the VNA-WALP depends on the adopted algorithm to solve (4.3). In order to compare different candidate algorithms, let us consider, as reference metrics, the number of iterations and function evaluations (i.e., data download/upload and measurement) needed to reach a maximum error across frequency $\|\overline{E}\|_\infty$ within a user-prescribed tolerance. With this choice, the results are largely independent from any particular technical WALP implementation, which will influence convergence speed just in terms of physical time. Four standard algorithms are evaluated in this work: Newton's method, nonlinear least-squares, Broyden's method and secants.

Newton's method

Newton's method is reported in literature as a suitable solution for active load pull for a low number of tones, such as the case of harmonic load pull [183], [192]. The algorithm uses the Jacobian matrix $\mathbf{J}$ of the partial derivatives of $\overline{E}$: $\mathbf{J}_{h,k}\left(\overline{a}_{S2}\right) = \left[\frac{\partial E_h}{\partial a_{S2}(f_k)}\right]$. The

solution at the r-th iteration is computed as:

$$\overline{A}_{S2}^{(r)} = \overline{A}_{S2}^{(r-1)} - \mathbf{J}_{h,k}\left(\overline{A}_{S2}^{(r-1)}\right)\overline{E}\left(\overline{A}_{S2}^{(r)}\right). \tag{4.4}$$

As the error function does not generally have an analytic expression, the partial derivatives have to be extracted using measurement-based finite-difference approximations. Each finite-difference implies that each of the N output frequencies is successively injected with a small single-tone signal excitation in order to compute part of the Jacobian. This greatly increases the number of required evaluations.

Nonlinear least-squares

The system in (4.3) can be recast as a nonlinear least-squares minimization, by searching the solution that minimizes the 2-norm of the measured error $\|\overline{E}\|_2$. While several gradient-descent or similar optimization algorithms are available for this task, here we focus on the trust-region algorithm of the built-in function *lsqnonlin* in MATLAB®. The computation of a numerical gradient still requires the same number of evaluations as the Newton's method.

Broyden's method

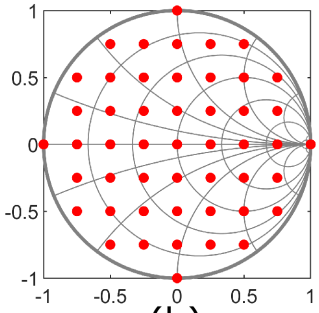
In order to reduce the large number of evaluations of Newton's method, Broyden's method computes the full $\mathbf{J}$ just at the first iteration, while $\mathbf{J}$ at the subsequent iterations are computed from the first one by adding rank-one matrix updates, which require a single measurement per iteration.

Secants' method

In the case of a perfectly linear DUT, any injected tone would influence the response just at the exact same frequency, as the absence of IM distortion prevents any regrowth or interaction among the different tones. Thus, $\mathbf{J}$ at each step would become diagonal, and each of the N frequencies could be treated separately using any method for finding zeroes of a 1-D function. The secants' algorithm is a well-known method [190] to efficiently perform this task. In the nonlinear case, where the $\mathbf{J}$ is full in general, the diagonal components are still expected to dominate the overall behaviour. Indeed typical PAs, when excited by a signal of a given frequency, will display a main response at the same frequency and a smaller, yet often relevant, nonlinear regrowth. In this context, the decoupling among the tones can still be considered a reasonable approximation. Any residual error introduced by this approximation, which in principle might severely influence the convergence, can be possibly compensated by an higher number of iterations. With respect to the previous three algorithms, the secants' method requires two distinct starting points instead of a single one. The first starting point is taken, as in other methods, as the trivial one

Algorithms	# FuncEvals
Newton	$P \times (2N+1)$
Lsqnonlin	$P \times (2N+1)$
Broyden	$(2N+1)+P$
Secants	$P+1$

(a)



(b)

Figure 4.3: (a) Required function evaluations for the considered algorithms after P iterations, for an N-tone input signal. (b) Load grid used for algorithm comparison.

in which no injection ($\overline{A}_{S2} = 0$) is applied. For the second one, we select a small-signal equal-amplitude random-phase tickle injection that covers all the N tones. In this way, the synthesized Γ_2 is slightly modified for all frequencies at the same time, so to jump-start the algorithm.

Method comparison

A comparison between different methods is required in order to evaluate the trade-off between accuracy and measurement speed that each one offers in the WALP scenario. Indeed, more accurate methods (based on more function evaluations) will potentially require less iterations before reaching convergence, thanks to a more thorough characterization of the DUT behaviour. Yet, the actual number of evaluations depends on the generally large number of tones N. Figure 4.3a summarizes the theoretical performance in terms of function evaluations per iteration for each of the considered algorithms, assuming that each one takes P iterations to converge on the N tones.
In order to study this trade-off, these four algorithms were compared on a $N = 7$-tone, 80 MHz-wide random-phase multitone signal at a frequency carrier of 1 GHz. The 7-tone signal was applied to the input of a packaged amplifier (Mini-Circuits ZFL11-AD+) and load pull was performed on the same frequencies, neglecting IM distortion. The target wideband load is specified as flat across the 7 frequencies in the 80-MHz BW, and its value is subsequently swept across the Smith Chart according to the benchmark shown in Fig. 4.3b. The maximum allowed error tolerance per tone for $\Gamma_T(f)$ is set to 0.01. The mean and standard deviation of the number of iterations required to reach such tolerance across all the synthesized load grid are then computed. The results are reported in Tab. 4.1 for different average power levels, ranging from deep back-off up to 3-dB compression, in order to evaluate the effect on the cross-frequency coupling due to DUT nonlinearity.
All the algorithms show similar performance in terms of number of iterations across the Smith Chart and for different power levels, whereas they require a significantly different

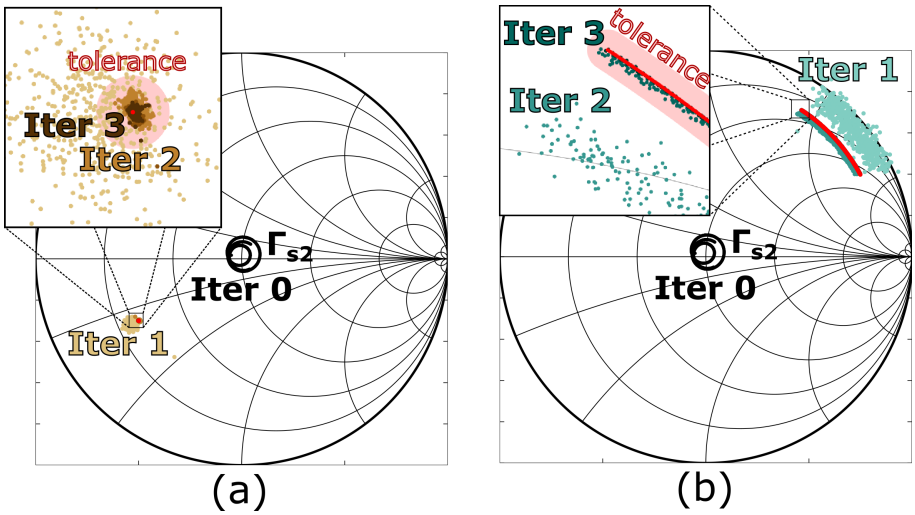


Figure 4.4: Iterative process to set the target load for two different profiles on the ZFL-11AD+ amplifier. Black at the 0^{th} iteration indicates no injected signal (starting condition, load equals the injection source match) and iterations use different shades of the same color. a) Fixed -0.5-j0.3 load across BW (brown). b) Fixed 0.6+j0.6 load seen through a $\frac{\lambda}{4}$ line at 1 GHz (turquoise). Red dots indicate the target load profile, while red shading highlights the 0.01 tolerance for convergence.

Table 4.1: Mean ± standard deviation across all the load grid for the number of iteration on ZFL-11AD+ PA. Excitation: 7-tone, 80 MHz signal at 1 GHz at different power levels.

P_{in} (dBm)	-27.2	-19.5	-12.2	-4.7
Newton	2.8±0.9	2.8±1.2	2.5±0.7	2.4±0.6
Lsqnonlin	3.2±1.5	2.94±0.85	2.8±0.7	2.7±0.7
Broyden	2.8±0.7	2.8±0.7	2.6±0.6	2.7±0.7
Secants	2.5±0.5	2.5±0.5	2.3±0.5	2.4±0.6

number of measurements before reaching the target. In this respect, the secants' method shows better performance in terms of convergence speed, given that the number of iterations is comparable to the other methods despite the distortion-induced mutual couplings among the tones, while the number of evaluations is significantly less. The independence of the number or evaluations from the number of tones in secants' case is particularly significant when considering broadband standard-like excitations, as multitones test signals with hundreds or thousands of tones are required to approximate the original standard to a good degree of accuracy [5], [64], [113], [173], [191].
The same favorable convergence behaviour was observed for the secants' methods on the DUT for random-phase multitones whose number of tones was increased from 7 up to

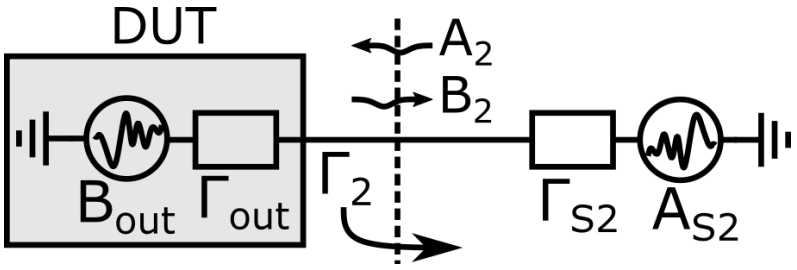


Figure 4.5: Linear model for the DUT and output injection source at the load reference plane.

533 in different tests (80 MHz with 150 kHz spacing). An example of the performance of the method in the case of $N = 533$ is shown in Fig. 4.4. Two different load profiles are shown (Load A and Load B), starting from the no-injected signal condition at the 0^{th} iteration and progressively reaching convergence with a tolerance of 0.01. These results strongly point to the secants' method as the most convenient candidate for the solution of the WALP equations in the case of a large number tones. Nevertheless, the framework described in the following is fully applicable to any other method that can efficiently estimate a solution to (4.3).

4.2.4 Device output match compensation

The main disadvantage shared by all methods considered in the previous section lays in their generality. In effect, those methods just try to iteratively find zeroes of (4.3), without using any knowledge of the actual functional relationship that links the injected signal with the synthesized load. In particular, the secants' method assumes 1) that the Jacobian $\mathbf{J}$ of the linearization is diagonally-dominant, and 2) that $\overline{E}$ can be locally linearized with respect to $\overline{A}_{S2}$.
The first hypothesis is likely satisfied in practical cases, considering the low-order nonlinearity displayed by typical DUTs (see Sec. 4.2.3), especially in the case of devices for communications applications. The second hypothesis is harder to verify, as the error function displays, in general, an unknown nonlinear relationship with the injected tones. Such a relationship can only be evaluated through direct DUT measurements, and despite the promising performance shown in Sec. 4.2.3, convergence of the secants' algorithm may fail altogether on certain DUTs. This happens if, at some intermediate iteration, the synthesized load strays too far from the target one, so that the local linearization of the error vs. injected tone relationship is not a suitable approximation, hindering the ability of the algorithm to converge to the required solution.
To further investigate this behaviour, let us consider a simplified linear model of the DUT within the WALP system, as shown in Fig. 4.5. Both the device and the output injection source can be represented as Norton/Thevenin equivalents (i.e., an ideal source and the

output match) at the load reference plane. In this case, the error at each frequency can be computed, using frequency domain waves, as:

$$E_n = \frac{\Gamma_{S2}(f_n) + \eta_n}{1 + \Gamma_{out}(f_n)\eta_n} - \Gamma_T(f_n); \quad \eta_n \stackrel{def}{=} \frac{A_{S2}(f_n)}{B_{out}(f_n)}; \tag{4.5}$$

where η_n is defined as the ratio between the output source injection $A_{S2}(f_n)$ and the injection $B_{out}(f_n)$ due to the DUT for a given frequency f_n.
Equation (4.5) can be obtained as follows. Referring to the schematic in Fig. 4.5, the Norton-Thevenin equivalent of the DUT output imposes the following relationship on the A_2 and B_2 traveling waves at each frequency f_n in the range of interest (with the explicit dependence omitted for clarity):

$$B_2 = B_{out} + \Gamma_{out} A_2. \tag{4.6}$$

Similarly, the effect of the output active load injection source can be described as:

$$A_2 = A_{S2} + \Gamma_{S2} B_2. \tag{4.7}$$

By solving the two equations (4.6) and (4.7) for the two unknown waves A_2 and B_2 results in

$$A_2 = \frac{A_{s2} + \Gamma_{s2} B_{out}}{1 - \Gamma_{s2}\Gamma_{out}}; \quad B_2 = \frac{B_{out} + \Gamma_{out} A_{s2}}{1 - \Gamma_{s2}\Gamma_{out}}. \tag{4.8}$$

Finally, the synthesized reflection coefficient Γ_2 can be found from the definition:

$$\begin{aligned} \Gamma_2 \stackrel{\text{def}}{=} \frac{A_2}{B_2} &= \frac{A_{s2} + \Gamma_{s2} B_{out}}{B_{out} + \Gamma_{out} A_{s2}} \\ &= \frac{\frac{A_{s2}}{B_{out}} + \Gamma_{s2}}{1 + \Gamma_{out}\frac{A_{s2}}{B_{out}}} = \frac{\eta + \Gamma_{s2}}{1 + \Gamma_{out}\eta} \end{aligned} \tag{4.9}$$

where η is defined as the ratio between the load injection source and the DUT output-equivalent source $\eta \stackrel{\text{def}}{=} \frac{A_{s2}}{B_{out}}$. This Möbius-type relationship predicts that, in the linear case, circles in the η-plane will map into circles in the Γ_2 plane. This means that injecting a certain output A_{S2} and sweeping its phase, the resulting Γ_2 will describe a circle across the Smith chart, with its radius and center depending on the source match, the output match and the mangitude of the injected signal. Conversely, the obtained functional relation also allows to predict the injection that is required in order to synthesize a given load reflection coefficient.
Equation (4.5) shows that, even in the case of a perfectly linear DUT, for which diagonality of **J** is trivial given the absence of IM, the error function at each frequency is nonlinear with respect to the injected signal A_{s2} at the same frequency. In particular, this happens when the DUT is strongly mismatched at the output ($\Gamma_{out} \neq 0$), which is the case for microwave

transistors, a typical target for load pull experiments. Indeed, the secants' algorithm, which locally linearizes the error function, will actually require multiple iterations, or might not converge altogether.
Equation (4.5) represents a closed-form expression for the error function. Hence, it can be analytically solved in order to find the value of η_n, and ultimately of $A_{S2}(f_n)$, to inject for synthesizing the required target at each selected frequency f_n. Eventually, once the DUT output match Γ_{out} and source match Γ_{S2} are known from measurements, it is possible to analytically solve the WALP problem in a single iteration. However, the validity of (4.5) is strictly limited to linear DUT operation, which is not a realistic case for microwave transistors for RF power applications. This, in principle, forbids the use of this method in the general case. Nevertheless, (4.5) can be used as a simplified model to exactly compute an initial approximate solution to the WALP problem. This solution, which takes into account the effect of the DUT mismatch, can actually provide a well-conditioned (i.e., close to the target) initialization to the secants' method.
First, let us assume that the injection source is operating linearly and independently from the DUT (due to the output circulator) at any power level. Then, in order to use the approximate linear model for a nonlinear DUT operating in LS conditions, we propose the use of a suitable generalization of Γ_{out} in order to find the correct injection signal. The output match compensation procedure consists of the following steps:

1. The DUT is excited at the input by a fixed multitone signal. A number N of tones at the output B_2 is selected for load pull at frequencies $f_1 \dots f_n$, which can include in-band spectral regrowth.

2. A small-signal (SS) random-phase multitone is concurrently injected by the output source at slightly offset frequencies. This output injection should be small enough to minimally perturb the LS operating point of the DUT, while at the same time providing good measurement dynamic range. This strategy results in two interleaved frequency grids displaying the separate results of the input or the output excitation. Γ_{S2} can be estimated on the input-excited frequencies, using the following ratioed VNA measurement at the output reference plane:

$$\Gamma_{S2}(f) = \left.\frac{A_2(f)}{B_2(f)}\right|_{@inEXC}. \tag{4.10}$$

   At the same time, a LS equivalent of the DUT output match can be estimated on the output-excited frequencies, again as a ratioed VNA measurement:

$$\Gamma_{out}(f) = \left.\frac{B_2(f)}{A_2(f)}\right|_{@outEXC}. \tag{4.11}$$

   The behavioral meaning of this LS output match parameter will be discussed in Sec. 4.2.6.

3. The measured Γ_{out} from (4.11) is smoothly interpolated across frequency, in order to provide estimation of its value at the input excited (i.e., non-offset) frequencies.

4. The output multitone tickler is rigidly shifted in frequency in order to align the two interleaved frequency grids. In this way, the measured $\Gamma_2(f)$ at the N target frequencies is tickled from the $\Gamma_{S2}(f)$ measured in (4.10), similarly to the secants' method starting point. This step allows to estimate the current injection ratio using (4.5) as:
$$\eta(f) = \frac{\Gamma_2(f_n) - \Gamma_{S2}(f_n)}{1 - \Gamma_{out}(f_n)\Gamma_2(f_n)}, \tag{4.12}$$
where the measurement of the reflection coefficient is used:
$$\Gamma_2(f) = \left.\frac{A_2(f)}{B_2(f)}\right|_{@in/outEXC}. \tag{4.13}$$

5. As a final step, the required injection ratio $\eta_T(f)$ to reach the target $\Gamma_T(f)$ can be computed at each of the N frequencies.
$$\eta_T(f) = \frac{\Gamma_T(f) - \Gamma_2(f)}{1 - \Gamma_{out}(f)\Gamma_T(f)} + \eta(f)\frac{1 - \Gamma_{out}(f)\Gamma_2(f)}{1 - \Gamma_{out}(f)\Gamma_T(f)}. \tag{4.14}$$
The ratio $\frac{\eta_T(f_n)}{\eta(f_n)}$ is then multiplied, frequency-by-frequency, to the numerical signal $A_{S2}(f_n)$ loaded in the output VSG in order to provide the required compensation:
$$A_{S2}^{comp}(f_n) = A_{S2}^{0}(f_n)\frac{\eta_T(f_n)}{\eta(f_n)}, \tag{4.15}$$

where $A_{S2}^{comp}(f_n)$ embeds the proposed compensation. When finally injected, $A_{S2}^{comp}(f_n)$ will synthesize the compensated Γ_{comp}.

4.2.5 Experimental measurement results

The proposed LS DUT output match procedure was tested on an un-matched Gallium Nitride HEMT (Macom NPBT00004A) biased in class AB. The input excitation is a random-phase multitone signal with 40 MHz BW around 5.8625 GHz and a tone spacing of 36 kHz, for a total of $N = 1111$ input-excited tones, with an available source power $P_{avs} = 13.7$ dBm. WALP is performed at the output reference plane on the same input-excited tones.
The load target profile Γ_T is set as a close approximation of the LS conjugate match of the DUT across the modulation BW, although this choice might not represent the actual optimal broadband impedance for best device operation. Indeed, as already mentioned in Sec. 4.1, finding such an optimum by sweeping all impedance combinations for a very

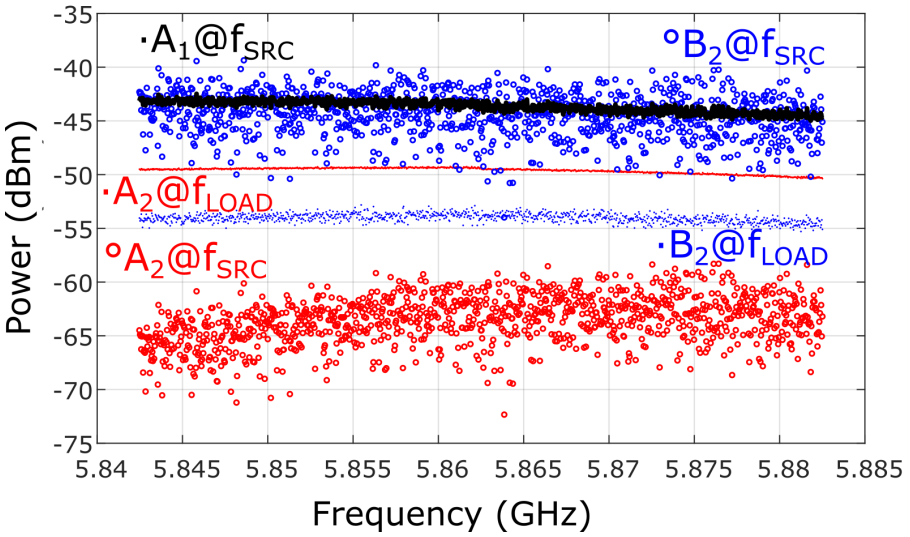


Figure 4.6: Interleaved excitation strategy for the estimation of Γ_{S2} and Γ_{out} on 40-MHz excitation BW for a GaN HEMT. A_1 (black), A_2 (red) and B_2 (blue) at source-side (circles) and load-side (dots) excited frequencies are reported.

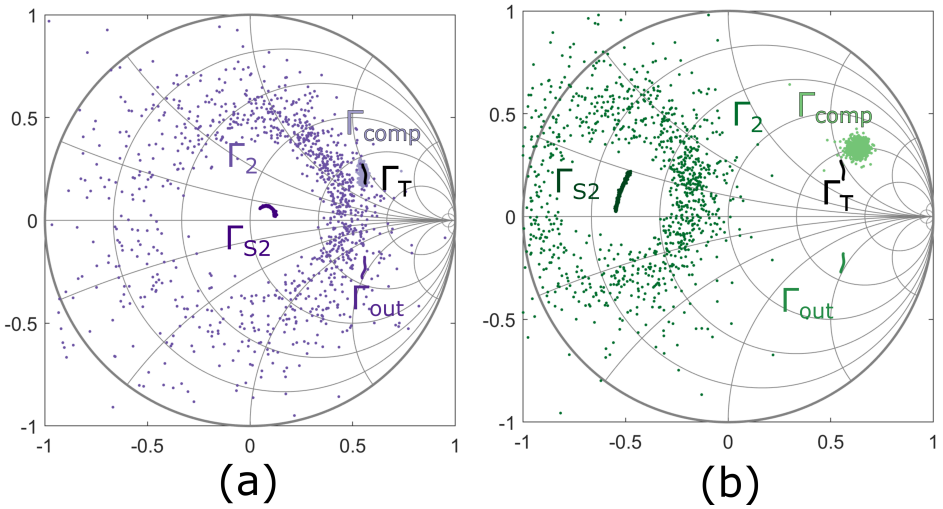


Figure 4.7: Reflection coefficients measured at different steps of the output match compensation procedure for two different output source matches. (a) Γ_{S2} obtained by the output PA and its circulator directly connected to the DUT. (b). Γ_{S2} synthesized using a passive tuner between the output source and DUT.

large number of frequency points across the BW, then producing load pull contours akin to the ones obtained for traditional narrowband load pull, is not practically feasible due to the high-dimensionality of the problem.

First, Γ_{S2} and Γ_{out} are measured on offset frequencies (step 2). Figure 4.6 shows the combined input-output excitations used in the second step of the compensation procedure. In particular, it can be seen that suitable ratioed measurements at interleaved frequencies can be used to separately estimate Γ_{S2} and Γ_{out} at the same time. Then Γ_2, which can be seen as the cloud of points representing a random *small* modification of Γ_{S2} at each tone, is measured on the original frequency grid (step 4). Finally, the compensated Γ_{comp}, which is reasonably close to the target Γ_T, is obtained by injecting the output match compensated signal (step 5). The reflection coefficients measured at the various steps of the compensation procedure can be observed in Fig. 4.7 for two different source matches Γ_{S2}. In the first case (Fig. 4.7a), the output PA with its circulator are directly connected to the DUT. In the second one (Fig. 4.7b), a passive tuner is added between the source and DUT (see Figs. 4.1 and 4.2) in order to emulate a load pre-match condition, such as the one found in hybrid passive-active load pull setups.
Moreover, the performance of the compensation procedure can be compared to the first iteration of the basic secants' method, in order to quantify the expected improvement for different loads. The tests are run for two different power levels (linear and compressed) of a single random-phase multitone signal with 237 tones across a 80 MHz BW at 5.85 GHz. The target load is taken to be a flat-across frequency profile, with the common value at all frequencies being swept across the Smith Chart. The RMS error, defined as

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |\Gamma_2(f_n) - \Gamma_T(f_n)|^2} \tag{4.16}$$

is used as a metric in order to compare the accuracy of the two techniques, provided that they are using the same number of measurements (i.e., two VNA frequency sweeps) after the first iteration.
Figure 4.8 reports the RMS error for the two methods and the two power levels as the target broadband load varies across the Smith Chart. The secants' method shows a significantly larger error at the first iteration, with the overall behaviour strongly depending on the target load due to the nonlinearity of the error function. As also shown in Fig. 4.4, the actual synthesized load Γ_2 can be outside of the Smith chart for some or all of the frequencies, possibly intercepting unstable load regions for the device. Instead, the linear output match compensation shows a significantly smaller RMS error (in most cases less than 0.1) just after the first iteration. These trends remain similar at low (Fig. 4.8a) and high power (Fig. 4.8b), with both methods showing an increase of the error in this second condition.
These experimental tests show that the procedure is able to provide a good starting guess for the signal to be injected to reach the target load. The estimation error is slightly larger for the case in which the source match is at a greater distance from the target load. These results prove that the linear compensation procedure is able to correct for the effect

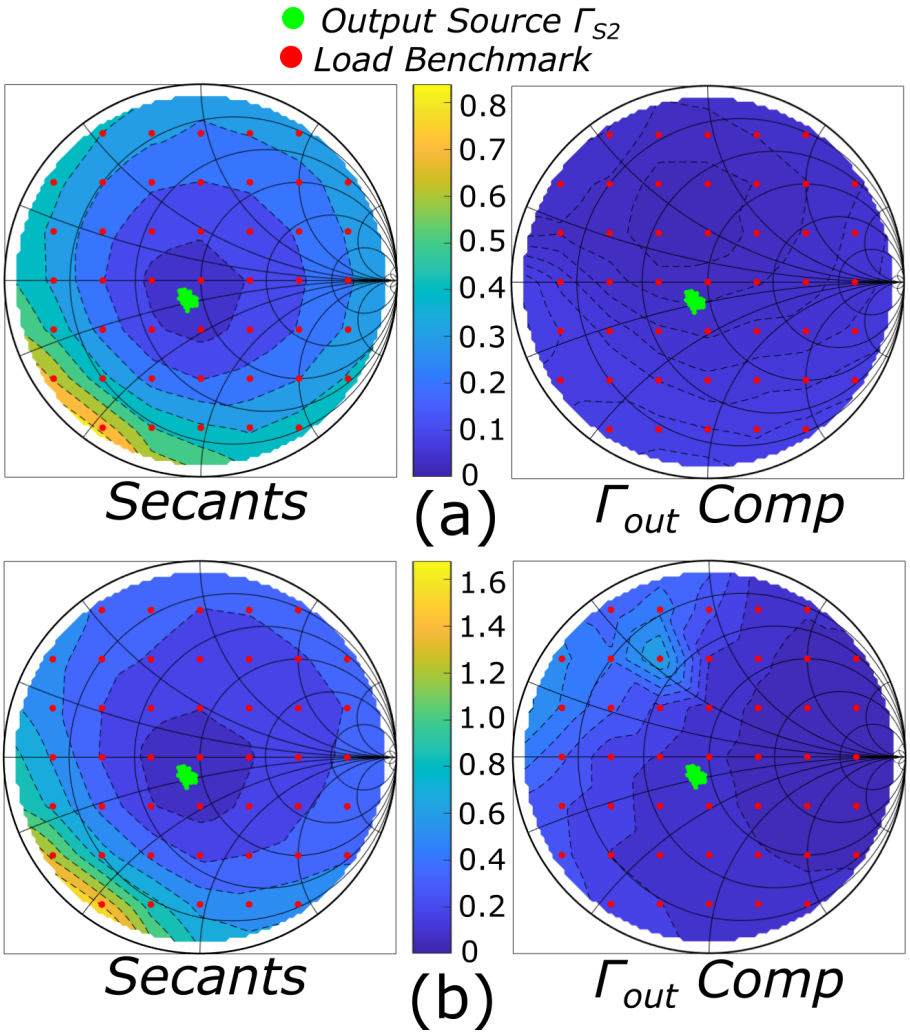


Figure 4.8: Comparison of rms error between first iteration of secants (left) and linear output match compensation (right) of Macom NPBT00004A transistor, for two different power levels. 237 tones at 5.85 GHz, for a total bandwidth of 80 MHz. a) Input power at 0 dBm. b) Input power at 15 dBm. The load benchmark (red) and the Γ_{S2} (green) of the output source are also reported.

of the DUT mismatch using a minimal number of measurements, and that it is directly applicable to hybrid load pull scenarios.

4.2.6 Estimation of the Large-Signal Output Match

As discussed in the previous sections, the output match compensation procedure is crucial in ensuring fast and stable convergence to the required target. Its effectiveness depends

on the use of a sufficiently accurate estimate of the LS output match (Γ_{out}) of the device. In this work, such an estimate is measured as the same-frequency ratio between the B_2 and A_2 waves as from (4.11), at those frequencies where the device is tickled by a SS multitone at the output port.
Several other techniques have been proposed in literature to estimate an LS equivalent of linear network parameters. Indeed, the Γ_{out} identified under SS conditions might be insufficient to accurately describe the LS device behaviour. Practically, the DUT behaviour at the output reference plane can be characterized as a SS linearization around an LSOP set by the input multitone a_{S1}. In the X-parameters framework [97], [193], the device response under load pull is described by means of un-perturbated LSOP (i.e., without active injection) plus a superposition of direct X_S and IM X_T terms, which can be identified using a SS single-tone tickler injected in the output, and then swept across the BW of interest. The approach soon becomes unwieldy for a large number N of load pulled tones, since tickling just one single output frequency already generates N cross-frequency IM terms (X_T) in the response. While the use of these terms would give a mathematically correct linearization around the LSOP, the comprehensive estimation of all the parameters would require a prohibitive amount of orthogonal VNA measurements [97]. Instead, the best-linear approximation (BLA) framework [194] and the hot-S_{22} approach [195] still use a single-tone output tickler, but focus just on the same-frequency direct X_S terms. Such terms, while being a rough approximation of the device behaviour, are just N in total and can be estimated using a VNA.
The solution adopted in this work, as from Sec. 4.2.4, uses a further modification of the previous methods, exploiting a SS tickler multitone [111] instead of a single tone. Indeed, as N increases, the power-per-tone decreases linearly, while the noise power density remains the same. Therefore, such a tickler multitone should be designed to have sufficient RMS power with respect to the VNA measurement noise, yet small enough not to perturb the LSOP. At the same time, the multitone phase distribution can be optimized [196] to yield a sufficiently low time-domain peak for a given RMS power. Within the compensation procedure, the excitation can be conveniently generated using the output VSG already available in the WALP setup.
In the proposed approach, the direct (i.e., X_S) and IM (i.e., X_T) terms overlap in determining the DUT response at each frequency and, as such, they cannot be separated in any way. Therefore, the raw measured response will generally depend on the specific phase realization of the input and output excitation multitones [193]. However, under the hypothesis that the input excitation is uncorrelated over different frequencies (e.g., a random-phase multitone), the N different X_T IM terms sum with random-phases and statistically average out. Therefore, for a large number of tones N [113], the measured LS Γ_{out} converges to the X_S-term that would be measured with a single-tickler hot-S_{22}, or applying the X-parameters approach. The main advantage of the tickler multitone is that it avoids the long source LO tuning times required for a SS source frequency sweep, while providing similar results and requiring the same number of measurements as the single-

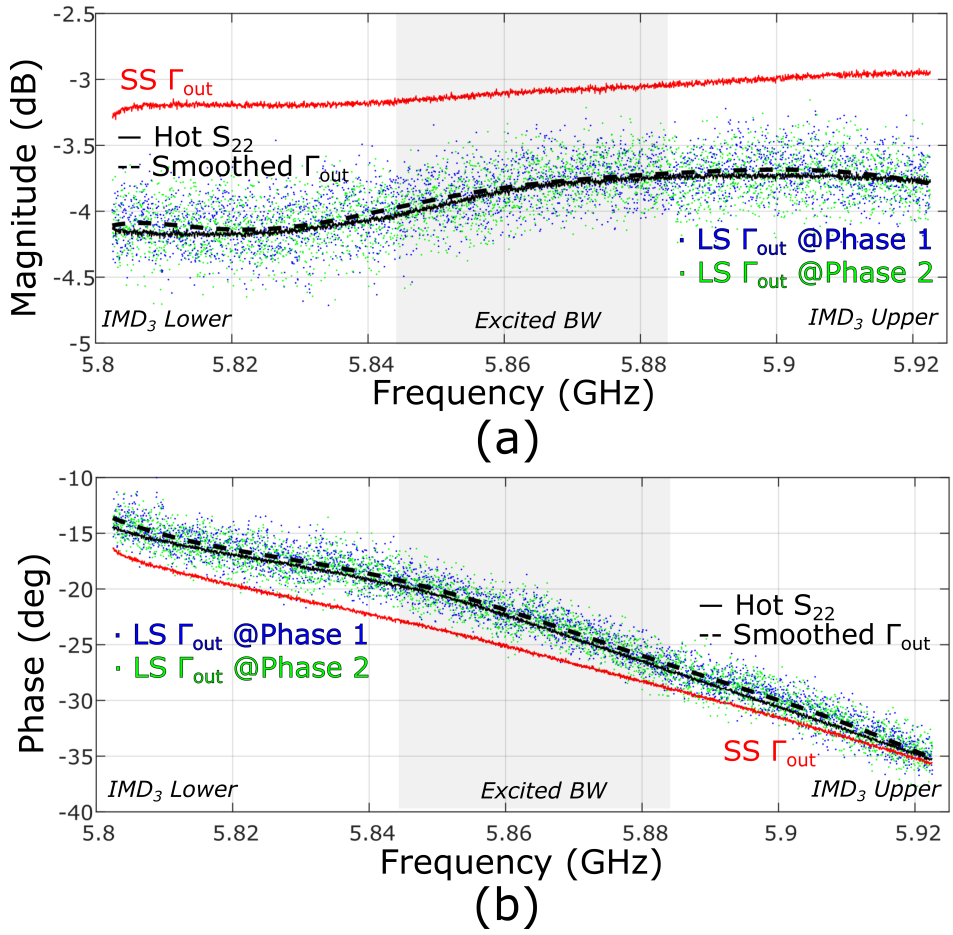


Figure 4.9: Comparison of the magnitude (a) and (b) phase of different definitions of the LS output match. SS Γ_{out} (red), hot-S_{22}-like Γ_{out} (dashed black) are compared with a smoothed version (solid black) of the multitone LS Γ_{out} measurement, as used in Sec. 4.2.4. Two multitone LS Γ_{out} measurements for two different (green and blue) phase realizations are reported. Grey shading identifies the input-excited 40-MHz BW and third-order IM BW.

tone approach. Indeed, it provides an instantaneous characterization across the whole BW of interest at the price of a reduced dynamic range, for a given total available tickle power. An experimental comparison obtained by applying the different definitions of the LS output match is reported in Fig. 4.9 for the GaN HEMT (biased in class-AB) introduced in the previous sections. First, the SS Γ_{out} is measured under linear operating conditions. Then, the DUT is excited using two different 40-MHz-BW LS random-phase multitones with $N = 1111$ tones and $P_{avs} = 13.7$ dBm, which will feature two different sets of phases realizations. The device is tickled on the output with a SS multitone covering both

upper and lower IM3 BWs (for a total of 120 MHz of output BW). This, similarly to the out-of-band-BLA [194], allows to estimate Γ_{out} at the IM3 frequencies.
The multitone LS Γ_{out} presents LSOP-dependendent, *noise-like* nonlinear stochastic distortions [111], [112] due to the overlap of direct and IM components. The value of the LS Γ_{out} is shown to depend on the actual phase realization of the multitone, and it is significantly different from its SS value. Nevertheless, the statistical uncorrelatedness of the different frequencies in the input signal allows to average the LS Γ_{out} measured with the multitone method across adjacent bins [113]. In this way, the smoothed profile (referred to as smooth Γ_{out} in Fig. 4.9) represents a statistical average that theoretically converges to an estimate of the X_S parameter, justifying the third step in the compensation procedure of Sec. 4.2.4. Finally, the hot-Γ_{out} (i.e., hot-S_{22}, equivalent to the X_S term) is measured by successively exciting the output of the DUT with a SS single-tone tickler swept across the BW of interest. As it can be observed, its value is extremely close to the smoothed LS Γ_{out}, experimentally confirming the theoretical analysis.
In conclusion, given the experimental results of Sec. 4.2.5 and the ones shown in [185], the LS output match measured using the multitone method is a suitable approximation to provide a first-step estimate of the solution to the WALP problem. On the other hand, more accurate models to describe the DUT LS output match behaviour could be adopted [113], possibly improving the pre-compensation. However, their experimental identification will typically require more extensive measurement capabilities, or a greatly increased number of acquisitions and measurement time. Indeed, in the WALP scenario, the goal is to achieve convergence in a reduced number of acquisitions/iterations, and not to build an extensive model of the DUT. In this respect, the proposed output match compensation provides a reasonable approximation that is fit for the purpose.
The approach is able to describe and compensate the variation of DUT output characteristics across frequency. Indeed, it constitutes the best (in the least squares sense [112]) linear time-invariant approximation of the output match for the given LSOP, providing a characterization of the frequency dynamics across the full BW of interest. Therefore, it expected that the compensation method will still provide the favorable experimental performance reported in this section also when arbitrarily wider BWs are considered.

4.2.7 Hybrid method performance

As pointed out in the previous Section, the output match compensation procedure can be effectively used to initialize the secants' method in a well-conditioned way. The two starting points for the algorithm are then taken to be the zero-injection case (in which $\Gamma_2 = \Gamma_{S2}$), and the injection resulting from the compensation procedure in Sec. 4.2.4. This results in a combined *hybrid method*, in which the first iteration estimates an approximate model of the DUT, while the following ones use the secants' method to blindly find the zero of the error function, yet in a restricted region close to the final target.

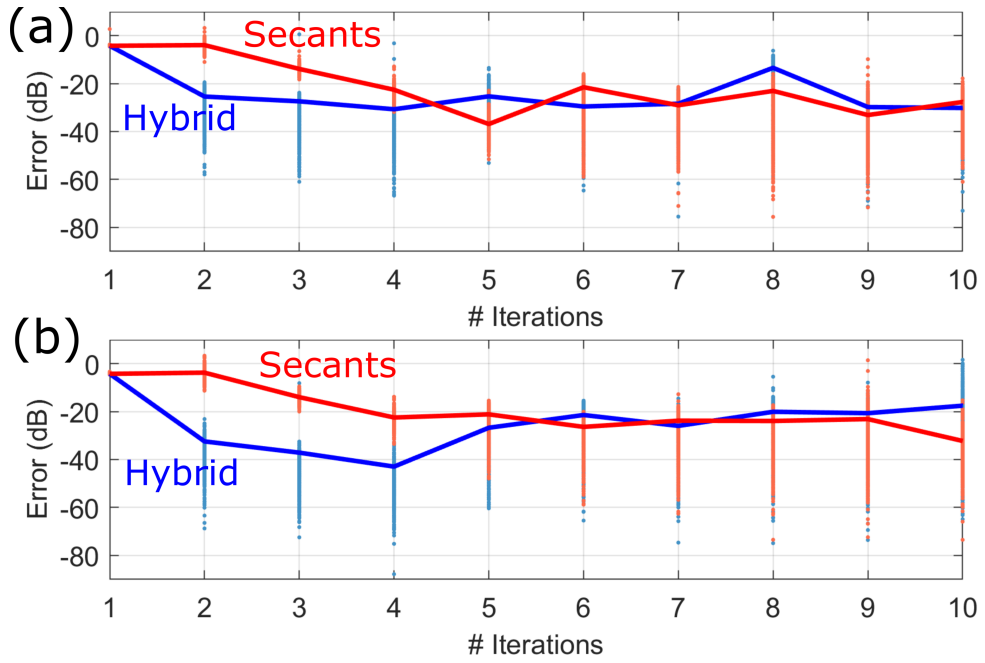


Figure 4.10: Comparison of the secants' (red) and hybrid (blue) method in the WALP of a 1111-tone 40-MHz wide multitone at 5.625 GHz. The two methods are both run for 10 iterations, and the frequency-by-frequency error (dots) and the RMS error (solid line) are reported.

In order to compare the performance of the standard secants' method with the proposed hybrid one, the GaN HEMT under test is excited using two different 40-MHz-BW random-phase multitones with $N = 1111$ tones and $P_{avs} = 13.7$ dBm, with the load target set as the conjugate to the LS output match in Fig. 4.9. The two algorithms are run for 10 iterations without imposing any other stopping criterion in order to evaluate the intrinsic performance of each method. Figures 4.10a-b report the comparison of the error for the two different input phase realizations using the same target load. It is possible to observe that, in both cases, the initialization provided by the compensation method allows the hybrid algorithm to converge much faster than the secants one, reaching a minimum RMS error of -30/-40 dB in 4 iterations. Nevertheless the secants' algorithm can reach the same precision if more iterations are allowed, 5 in the first case and up to 10 in the second one. Once the minimum is reached, the error in the hybrid method starts to rise with the successive iterations, which are computed using the blind secants' algorithm, and in effect display a performance similar to the unmodified secants' method. This paradoxical behaviour is due to the fact that, once a low enough error is achieved, the variations in the measured reflection coefficients from one iteration to the next are comparable with the measurement noise, and the iteration becomes ill-conditioned. In other words, the error does not decrease monotonically with the iterations before reaching a plateau dictated by

noise. Therefore, it is not merely sufficient to wait a large enough number of iterations before reaching any given precision. It is instead of utmost importance to find a suitable stopping tolerance that allows to reach a small error, but avoids ill-conditioning once the error hits the setup noise floor (see Sec. 4.2.8).
Still, the proposed technique allows to pick the minimum error solution among the synthesized ones, obtaining the user-prescribed impedance within the tolerance allowed by noise. Actually, as can be seen in Fig. 4.10, the linear compensation step used by the hybrid method is fundamental to greatly improve the conditioning of the iterative procedure, by preliminarily setting the active injection in the neighbourhood of the final solution. In this sense, the hybrid method enables greater precision in setting the user-prescribed impedance within a limited number of iterations.
For both cases in Fig. 4.10, it can be noticed that, even if the RMS error is sufficiently low, there might be some frequencies at which the error is still quite noticeable (i.e., $\|\overline{E}\|_\infty$ defined in Sec. 4.2.3 is above the tolerance). However, these few outliers do not compromise the DUT characterization, as the LSOP is set by the reflection coefficient seen at each of the $N = 1111$ tones. In practice, each specific tone has then only a minor influence on the overall DUT behaviour. Moreover, the use of these methods allows to start and stop the iteration separately at each frequency. Therefore, it is possible to stop the iteration for the tones where sufficient accuracy is reached and continue iterating on the tones which present a reflection coefficient outside the tolerance. The effect of cross-frequency IM coupling can be accounted for by dynamically starting and stopping iterations at each tone. The implementation and testing of this dynamic iteration and stopping procedure lies beneath the scope of this work.

4.2.8 Dynamic Range of the VNA-WALP

The VNA-WALP functionality critically leverages on the availability of accurate measurements of the synthesized output reflection coefficient $\Gamma_2(f)$ at each iteration in order to compute the error function. Indeed, if the measurements are too noisy, any iterative algorithm can only reach a rough approximation of the required target, or it might diverge altogether. Since $\Gamma_2(f)$ is defined as the ratio between the $A_2(f)$ and $B_2(f)$ waves, its value might be ill-conditioned at the frequencies for which $B_2(f)$ is close to the noise floor (i.e., the DUT has almost-zero output). Conversely, this issue is not present in WALP setups based on full-waveform measurement capabilities [75], [85], [184]. Indeed, as described in Sec. 4.2.1, the iterative method in those cases uses waveform measurements and multiplication by the target load, without having to estimate ratioes at any step for reaching convergence. Moreover, in the VNA-WALP approach, the dynamic range of the setup cannot be artificially increased in post-processing beyond the noise floor, e.g., by using smoothing methods [191]. Indeed, actively synthesized reflection coefficients $\Gamma_2(f)$ at each iteration cannot be assumed to be smooth across frequency, even if the targeted reflection coefficient is. Nevertheless, these drawbacks are largely mitigated by the raw

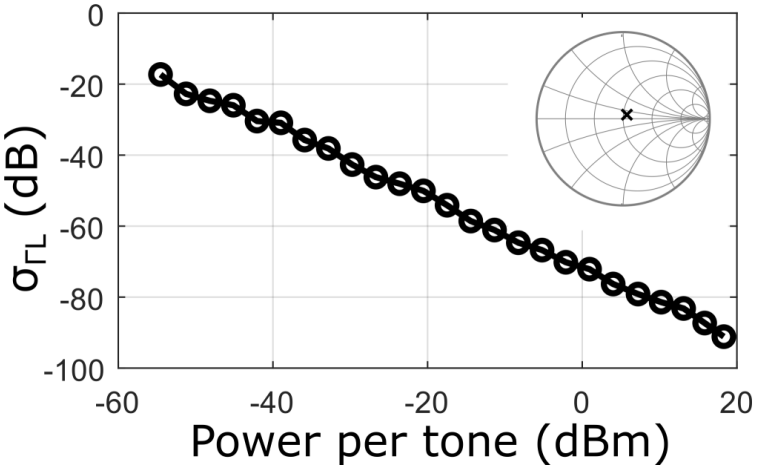


Figure 4.11: Measurement noise standard deviation in function of power per tone at 5.625 GHz. The load reflection, in the Smith Chart inset, is the one seen by the DUT when active injection is turned off.

dynamic range of modern VNAs, which can be further increased by using narrower IF BWs and coherent averaging techniques, even at high operating frequencies. Moreover, it can be reasonably expected that the exact value of the load reflection coefficient at frequencies with low available ouput power will barely contribute to the determination of the device LSOP, thus still allowing for a reasonable DUT characterization.
In order to evaluate the intrinsic dynamic range of the setup, the output coupler (Fig. 4.1) was excited by directly connecting the input VSG, which was set to synthesize a single-tone excitation at 5.635 GHz. No active load pull injection was applied, resulting in the passive load presented by the output tuner and circulator. This passive $\Gamma_2(f)$ was measured 25 times at different power levels. At each power level, the standard deviation across the repeated measurements was computed, as reported in Fig. 4.11. As expected, the empirical uncertainty in the Γ_2 due to measurement noise decreases with increasing available power at the output reference plane. This single tone characterization is sufficient to characterize the general multitone case, as the measurement test-set is assumed to operate linearly (i.e., superposition is valid) across all the power range. In any case, the results might depend on the specific microwave test-set (i.e., couplers and attenuators), and possibly, on the measured load [197]. Figure 4.11 can be used to measure the maximum accuracy in the load measurements, and to provide an estimate of the expected tolerance for a given power. This circumvents possible pathological behavior of the iterative method by avoiding further computations once the noise floor is reached (Sec. 4.2.3). Assuming a gaussian distribution for the noise, a confidence interval of $\pm 3\sigma$ (i.e., 9.5 dB in excess of the curve in Fig. 4.11) seems to be a resonable range to set the maximum expected accuracy. In addition, the estimation of dynamic range allows to gauge a suitable power

level for the output tickler multitones used in both the secants' and the hybrid method. Such ticklers must be small enough not to perturb excessively the LSOP of the device (which may hamper the convergence of the iteration) but, at the same time, to provide sufficient dynamic range in ratioed measurements.

4.2.9 Considerations on 5G-OFDM-like signals

These dynamic range considerations are particularly relevant when the DUT excitation signal is composed by a large number of tones. Hundreds or thousands of tones [112],[113] are typically required to provide an accurate spectral and statistical approximation of 5G-OFDM standards, and to set a realistic LSOP for the DUT. The power of the $B_2(f)$ wave will then spread across all the output tones, including IMs. For example, the in-band power-per-tone in the previously examined random-phase multitone with $N = 1111$ will roughly be 30 dB less than the RMS value of the output power, while the value at IMs frequencies will be considerably less.
This behaviour is shown in Fig. 4.12 for the GaN HEMT under test when the input is excited by 40-MHz-wide random phase multitones. The spectrum of a specific realization and the power spectral density (PSD) estimate (over 25 realizations) of the underlying stochastic process are reported for the $A_2(f)$ and $B_2(f)$ waves at the DUT output. It can be noted that, even if the PSD is relatively smooth and presents a well-defined power level at each frequency, a single realization can present large statistical variations around that level for the $B_2(f)$ and $A_2(f)$ signals, despite the flat-amplitude injection at the input. In particular, dips of up to 15 dB below the the power level of the $B_2(f)$ PSD can be observed in the upper and lower IM3 BWs. As any WALP algorithm works on a specific periodic realization of the process (and not on the statistical PSDs), this analysis further highlights the need for sufficient dynamic range in the Γ_2 across all the measurement BW. The situation is possibly more problematic for compact test [113] or other gaussian signals [112], [191]. In contrast to flat-amplitude multitones, such excitations present noise-like characteristics even in the excited band, where sharp dips in the A_1 (and B_2) amplitude can be present even in excited in-band components for a given realization. In this light, random-phase flat-amplitude multitones, which can likewise be tailored to properly approximate 5G signals [112], [173], seem to have a definite advantage in terms of dynamic range.
As the number N of tones has been shown to be the critical variable in determining the dynamic range requirements of the WALP setup, a potentially problematic situation could arise when measuring extremely wide BWs (such as the ones enabled by the proposed VNA-based approach) with narrow tone spacings. Then, the power-per-tone at the output of the device would be extremely reduced for a given output power, with the resulting low signal-to-noise ratio in the load measurements, hence poor stability of the proposed methods. The solution that is typically adopted [113] is to fix a number N of tones, typically in the thousands range, that will allow to match the ccdf of the application

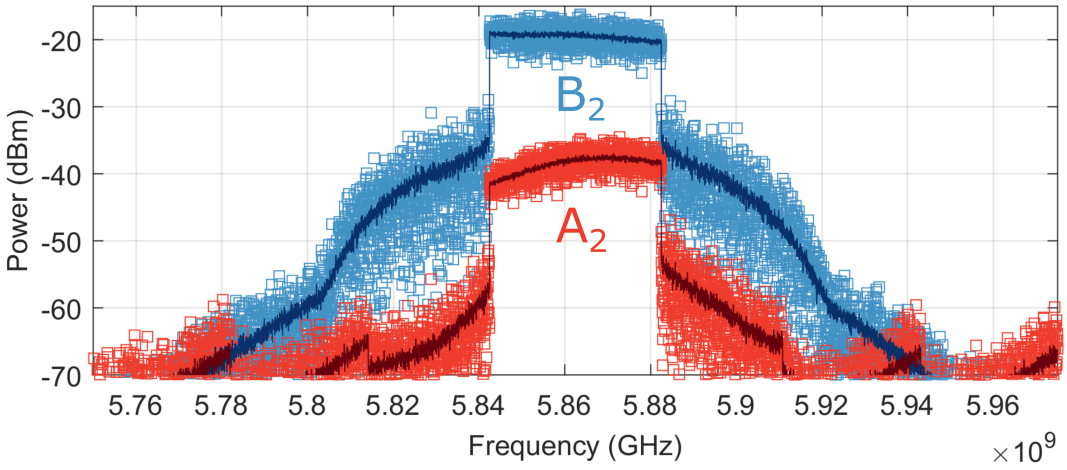


Figure 4.12: PSD across frequency for the A_2 (red) and B_2 (blue) waves for the GaN HEMT under test when excited by a $N = 1111$ flat-amplitude random-phase multitone signals. Darker continuous lines represent a statistical average of the PSD of the stochastic processes, while lighter squares show the particular realization.

signal with a good degree of accuracy. Then, the tone spacing is adjusted accordingly in order to cover the wide modulation BW of interest without reducing the power-per-tone below the critical level shown in Fig. 4.11. In some applications, however, the tone spacing might influence the long-term memory behaviour of the DUT [173] and cannot be set freely by the user. In that case, the dynamic range of the setup has to be extended either using lower noise HW components or by reducing the IF BW and exploiting coherent averaging techniques.

4.2.10 VNA-WALP at out-of-band frequencies

Given that sufficient measurement accuracy is available, the VNA-WALP framework can be applied to perform load pull even at out-of-band frequencies. An example of the achievable error by the current VNA-WALP setup is shown in Fig. 4.13. The input excitation is a 40-MHz-BW random-phase multitone on the input ($N = 1111$), with an available source power of $P_{avs} = 18$ dBm. VNA-WALP is performed across the full 120 MHz IM3 BW, with the load target set as the conjugate to the LS output match as shown in Fig. 4.9. It can be observed that the reached accuracy is significantly higher in the input-excited band and decreases across the IM3 BW, as the $B_2(f)$ power and dynamic range progressively drop.

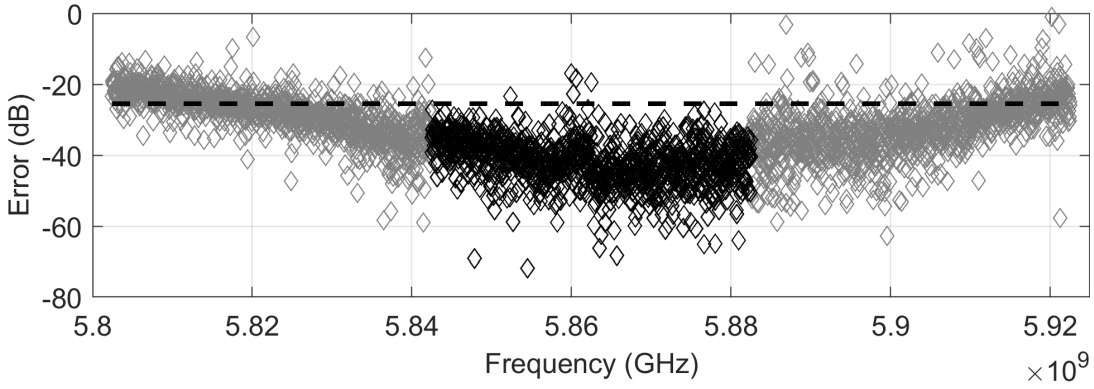


Figure 4.13: Error across frequency for a 40-MHZ input random-phase excitation and a 120 MHz VNA-WALP across the third order output BW. Input excited BW (black), third-order IM BW (grey) and RMS error across frequency (dashed line) are reported.

4.3 EVM and load pull

4.3.1 Measurement-based EVM model

In the following discussion, the measurement-based EVM framework introduced in Sec. 1.4.3 will be suitably generalized in order to extend its descriptive power to model a microwave DUT under load pull conditions.
Using the BLA decomposition shown in Fig. 4.14a, for a generic input spectrum $X(f)$, the output spectrum $Y(f)$ for any SISO nonlinear-dynamic period-preserving system can be expressed as:

$$Y(f) = G_{YX}(f)X(f) + D_{YX}(f) + N_Y(f) = \overline{Y}(f) + N_Y(f); \tag{4.17}$$

where G_{YX} is the least-square best approximation of a linear dynamic frequency response function (FRF) of the system across all the excitations for a given signals' class [63]; $D_{YX}(f)$, a zero-mean contribution with variance $\sigma^2_{D_{YX}}(f)$, quantifies the input-uncorrelated nonlinear stochastic distortions; $N_Y(f)$, with variance $\sigma^2_{N_Y}(f)$, represents additive measurement noise.
Using (4.17), the EVM_{YX} results [116]:

$$\mathrm{EVM}_{YX} = \sqrt{\frac{\int_{\mathrm{BW}} \sigma^2_{D_{YX}}(f)df}{\int_{\mathrm{BW}} |G_{YX}(f)|^2 S_{XX}(f)df}} \tag{4.18}$$

where $S_{XX}(f)$ is the input power spectral density (PSD), the expression $|G_{YX}(f)|^2 S_{XX}(f)$ depicts the linearly input-correlated PSD across the excitation BW, and

$$\sigma^2_{D_{YX}}(f) = S_{YY}(f) - |G_{YX}(f)|^2 S_{XX}(f) \tag{4.19}$$

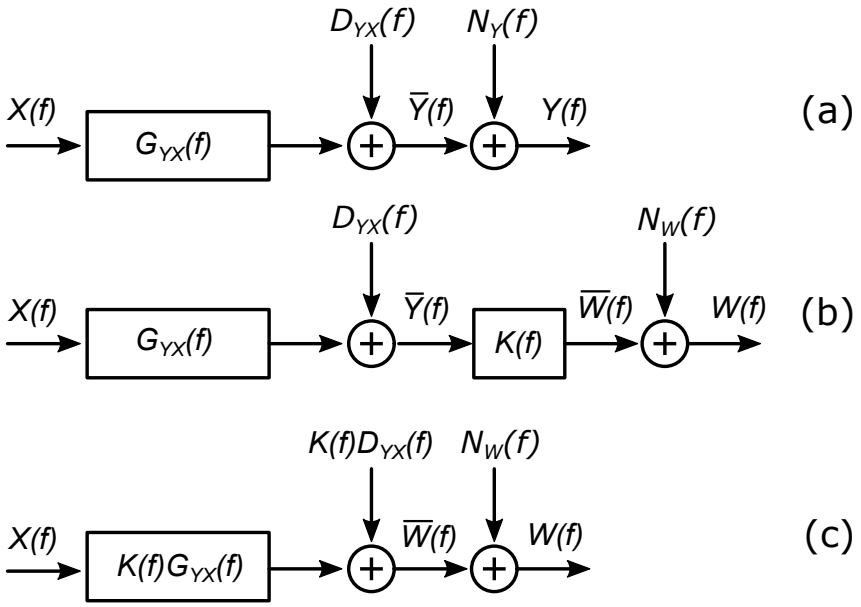


Figure 4.14: Measurement-based EVM model. (a) Case for a nonlinear-dynamic system. (b) Case for a nonlinear-dynamic system cascaded with a linear system $K(f)$. (c) Case for a nonlinear-dynamic system corresponding to (b).

is the purely nonlinear distortion power, $S_{YY}(f)$ being the output PSD.
In this section, we extend this basic framework by studying the effect on the distortion that can be observed by cascading the PISPO system of interest with a linear time-invariant network $K(f)$. In the application of interest, K will be related to the load termination seen by the DUT, but this description is general enough to cover other cases as well.
The general situation is represented by the block diagram in Fig. 4.14b. The combination of the two networks is again a PISPO system, with its new output W described as:

$$W(f) = G_{WX}X(f) + D_{WX}(f) + N_W(f) = \overline{W}(f) + N_W(f). \tag{4.20}$$

By neglecting output measurement noise, which is independent from all other quantities, the intermediate variable $\overline{W}$ representing the BLA response plus the nonlinear stochastic distortions can be expressed using (4.17):

$$\overline{W(f)} = G_{WX}X(f) + D_{WX}(f) = K(f)\overline{Y}(f) = K(f)G_{YX}X(f) + K(f)D_{YX}(f). \tag{4.21}$$

Therefore, by comparing the two terms, the following relationships hold

$$\begin{aligned} G_{WX} &= K(f)G_{YX} \\ D_{WX}(f) &= K(f)D_{YX}(f). \end{aligned} \tag{4.22}$$

This result implies that the cascaded linear gain multiplies both the BLA fo the original PISPO system and its nonlinear stochastic distortions. Finally, the overall EVM is then

calculated as:

$$\mathrm{EVM}_{WX} = \sqrt{\frac{\int_{\mathrm{BW}} \sigma^2_{D_{WX}}(f)df}{\int_{\mathrm{BW}} |G_{WX}(f)|^2 S_{XX}(f)df}} = \sqrt{\frac{\int_{\mathrm{BW}} |K(f)|^2 \sigma^2_{D_{YX}}(f)df}{\int_{\mathrm{BW}} |K(f)|^2 |G_{YX}(f)|^2 S_{XX}(f)df}}. \quad (4.23)$$

From (4.23), it can generally be seen that $\mathrm{EVM}_{WX} \neq \mathrm{EVM}_{YX}$, while an exact equality holds if the amplitude response of the LTI network is flat across frequency, that is $|K(f)| = K \ \forall f \in \mathrm{BW}$. More in general, by considering the upper and lower bound of $|K(f)|^2$ in the frequency interval of interest in (4.23), it can be shown that:

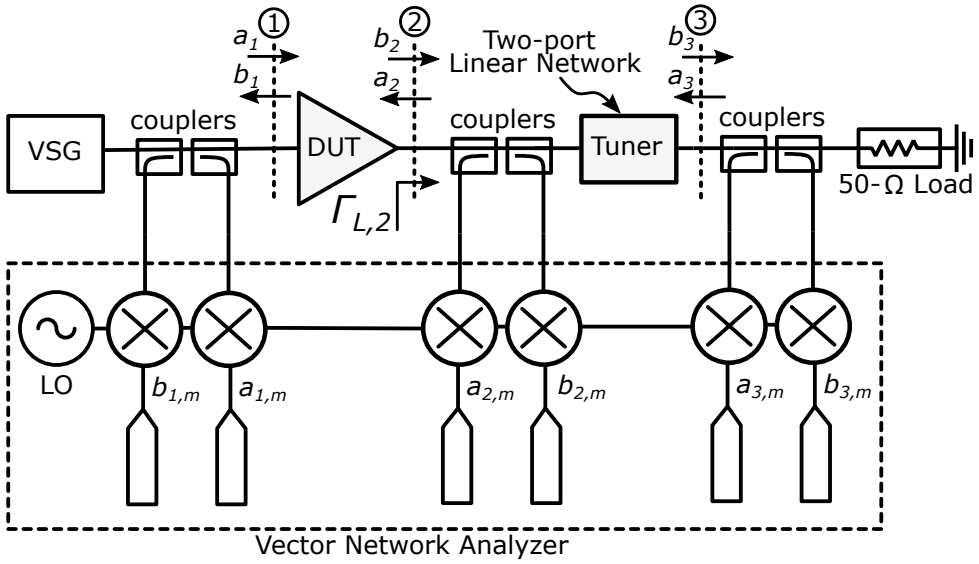
$$\sqrt{\frac{\min\limits_{f\in\mathrm{BW}} |K(f)|^2}{\max\limits_{f\in\mathrm{BW}} |K(f)|^2}} \mathrm{EVM}_{YX} \leq \mathrm{EVM}_{WX} \leq \sqrt{\frac{\max\limits_{f\in\mathrm{BW}} |K(f)|^2}{\min\limits_{f\in\mathrm{BW}} |K(f)|^2}} \mathrm{EVM}_{YX}. \quad (4.24)$$

4.3.2 Measurement setup

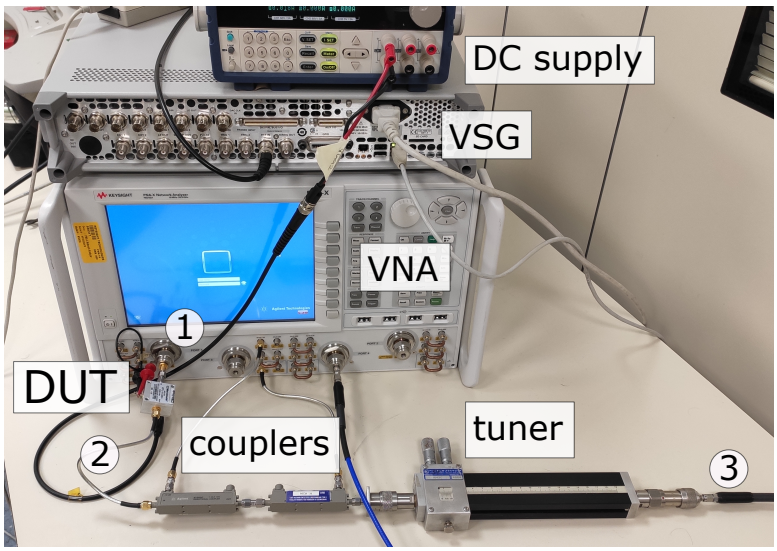
The main goal of the proposed characterization is to analyze the EVM behavior when different traveling waves on a given DUT are considered as input (i.e., X) or output (i.e., Y and W) signals, similarly to what is observed in load pull applications. In this situation, all the quantities of interest can be measured using random-phase multitone excitations and power-calibrated VNA-acquisitions, as described in Sec. 1.4.3.
The VNA-based measurement setup used in this work is reported in Fig. 4.15. The DUT is an off-the-shelf PA from Mini-Circuits excited with a 100-MHz-BW, L-tone random phase multitone signal (L =1001) centered at 1.75 GHz. The frequency spacing among the tones is 100 kHz. The signal PDF is gaussian with band-limited white PSD, reproducing a suitable test signal as from the 5G FR1 standard [3]. This input signal is generated by a Keysight MXG N5182B VSG, while the VNA in use is the Keysight PNA-X N5242A.
The DUT is cascaded with a manual slide screw tuner (Maury Microwave 7941A), which acts as tunable two-port linear network described by the following general transfer matrix:

$$\mathbf{T}_{32}(f) = \begin{bmatrix} t^{11}_{32}(f) & t^{12}_{32}(f) \\ t^{21}_{32}(f) & t^{22}_{32}(f) \end{bmatrix}. \quad (4.25)$$

Three ports and corresponding calibration reference planes are defined, each port using a two directional couplers (or reflectometers) for wave sensing (internal VNA couplers for ports 1 and 3, external couplers for port 2). The first two ports correspond to the input and the output of the PA, while the third one lays after the tuner. The VSG is connected to a rear-panel connector of the VNA and internally re-directed to inject into port 1. A classic three-port short-open-load-thru relative calibration has been performed, as well as a receiver power calibration referenced to a power meter. All six waves at the three ports are measured in the experimental tests.



(a)



(b)

Figure 4.15: Block diagram (a) and photo (b) of the VNA-based measurement setup used for EVM characterization.

4.3.3 EVM of PA under load mismatch conditions

The proposed configuration in Fig. 4.15 allows to analyze some fundamental aspects of modulation distortion measurements, for which two relevant cases are of interest:

(A) The distortion performance of a PA is known to be strongly influenced by the load

[75], as the LSOP is in general different for each termination. Picking a fixed load, is the measured distortion on the A_2 and B_2 wave the same? Which is the most suitable wave on which EVM performance has to be evaluated?

(B) The use of the tuner emulates the behavior of the output matching network in a PA, which adapts the optimal termination for the PA to the input impedance of the subsequent stages (e.g., antenna, 50 Ω line or other stages). Is the measured EVM the same before and after this two-port linear matching network (i.e., at reference planes 2 and 3, respectively)? Indeed, if the measured EVM was different between the two cases, a wideband load pull characterization would be not sufficient in order to fully predict the overall distortion in the PA.

These questions can be addressed using the framework in Sec. 4.3.1.
In this context, it is worth analyzing if $\mathrm{EVM}_{B_2A_1}$, i.e., with $B_2(f)$ as output and $A_1(f)$ as input of the DUT, is an effective metric in representing the modulation distortion of the whole PA-tuner chain. Indeed, it can be argued that the distortion measured on the reflected wave $A_2(f)$ or on the waves at the load reference plane at port 3 might be different, as from (4.23). This case can be analyzed by measuring $\mathrm{EVM}_{A_2A_1}$ and $\mathrm{EVM}_{B_3A_1}$. As a reference, it is know from theory that $\mathrm{EVM}_{B_2A_2} = 0$, as the relation between the two quantities is in general linear, as from (4.26).
The general framework can be specialized by taking the DUT input-output variables, as commonly reported in literature [99], $Y(f) = B_2(f)$ and $X(f) = A_1(f)$. On the other hand, the LTI gain $K(f)$ can have different espressions, depending on the variable of interest.
In case A, the cascaded linear two-port network is represented by the tuner input reflection coefficient that synthesizes a load mismatch for the PA, for which the $K(f) = \Gamma_{L,2}(f)$ and

$$\Gamma_{L,2}(f) = \frac{A_2(f)}{B_2(f)}. \tag{4.26}$$

Instead, for case B, the full relation imposed by the tuner between the waves at the two output reference planes is:

$$\begin{bmatrix} A_3(f) \\ B_3(f) \end{bmatrix} = \mathbf{T}_{32}(f) \begin{bmatrix} A_2(f) \\ B_2(f) \end{bmatrix}. \tag{4.27}$$

The typical case in which $Z_{L,3} = 50\ \Omega$ and $A_3(f) = 0$, that assumes that the tuner is used to match the PA to the reference impedance, resulting in

$$B_3(f) = (t_{32}^{21}(f)\Gamma_{L,2}(f) + t_{32}^{22}(f))B_2(f), \tag{4.28}$$

with $A_2(f) = \Gamma_{L,2}(f)B_2(f)$. Therefore, this case can be examined by taking $K(f) = t_{32}^{21}(f)\Gamma_{L,2}(f) + t_{32}^{22}(f)$.

While, theoretically, different EVM on different waves should have different values depending on $K(f)$ as from (4.20), in many load pull applications the constant $|K(f)|$ case might be the most interesting one. Indeed, the whole signal bandwidth is usually matched to the same practically-constant value [75], as electron devices typically display very broadband behavior in unmatched conditions. Electrical delays, such as those introduced by transmission line effects, share the same type of analysis, as the effect can be modeled by a rotation on the Smith-chart which just affects the phase of $K(f)$.
While non-flat-amplitude load profiles can be indeed emulated with WALP and implemented in MMICs, the use of these non-standard terminations is quite rare. Even in this case, the EVM differences might still be low, depending on the behavior-across frequency of the various terms in (4.20). So, in many load pull applications, it can be expected that a reasonable EVM estimation can be obtained independently from which specific signal is used for characterizations.
The reported analysis also clarifies why EVM is a "privileged" metric for characterizing distortion in RF PAs. For example, it could indeed be possible to select a linear matching $\mathbf{T}_{32}(f)$ so that $B_3(f)$ is equal to $B_2(f)$ in the bandwidth of interest and zero outside, possibly filtering out all out all the spectral regrowth components. Therefore, an ACPR measurement on B_2 would yield a finite value depending on the compression level, while B_3 would have and ACPR of zero. Therefore, this procedure might be incorrectly thought as a linearization of the PA, but it is only useful to avoid spectral leakage in adjacent channels, yet leaving the in-band distortion component untouched.
EVM, on the other hand, quantifies the in-band distortion, which shares the same frequency support as the refence signal to be amplified. Any cascaded passive LTI system can indeed reduce (or in general modify) EVM as in (4.23), but this also inevitably introduces modifications on the desired output signal at some level. In this respect, in-band distortion cannot be filtered out by any linear means and instead requires dedicated predistortion techniques.
Experimental results investigating the validity of these claims are reported in the text section.

4.3.4 Measurement results

The robust measurement procedure described in Sec. 1.4.3, with $P = 2$ and $M = 25$, has been applied for characterizing $\mathrm{EVM}_{B_2A_1}$ at different RF available source power $P_{av,s}$, obtaining the results in Figs. 4.16-4.18 for the $\Gamma'_{L,2}$ profile reported in the inset of Fig. 4.18.
The frequency-domain quantities in Fig. 4.16 are acquired with VNA frequency sweeps across the 100-MHz BW with receivers intermediate frequency BW of 1 kHz. The in-band distortion $D_{B2A1}(f)$ can be measured with a remarkably low noise level, showing a continuous profile with the out-of-band re-growth [113], [116]. The G_{B2A1} (Fig. 4.17) depends on the large-signal operating point (hence, on $P_{av,s}$ and load profile).

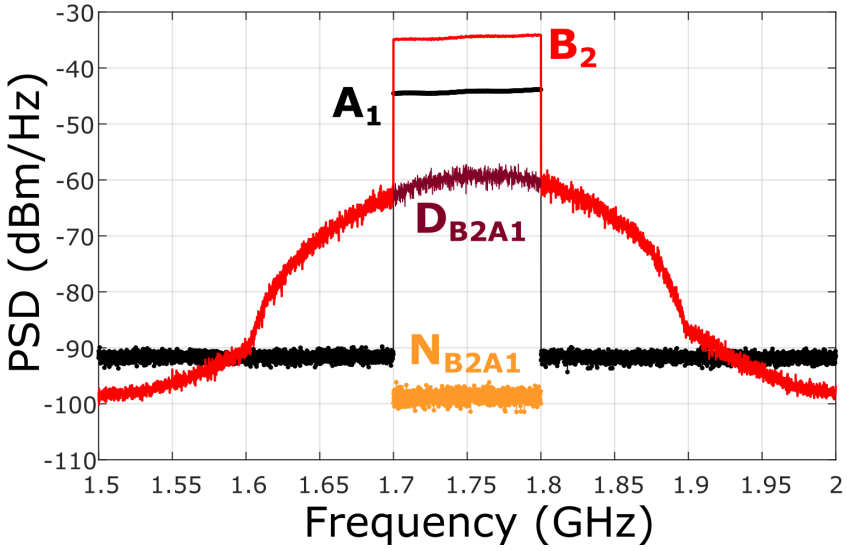


Figure 4.16: Frequency spectra of the quantities measured for characterizing $\mathrm{EVM}_{B_2A_1}$ ($\Gamma'_{L,2}(f)$ load, $P_{av,s}$=-6.9 dBm).

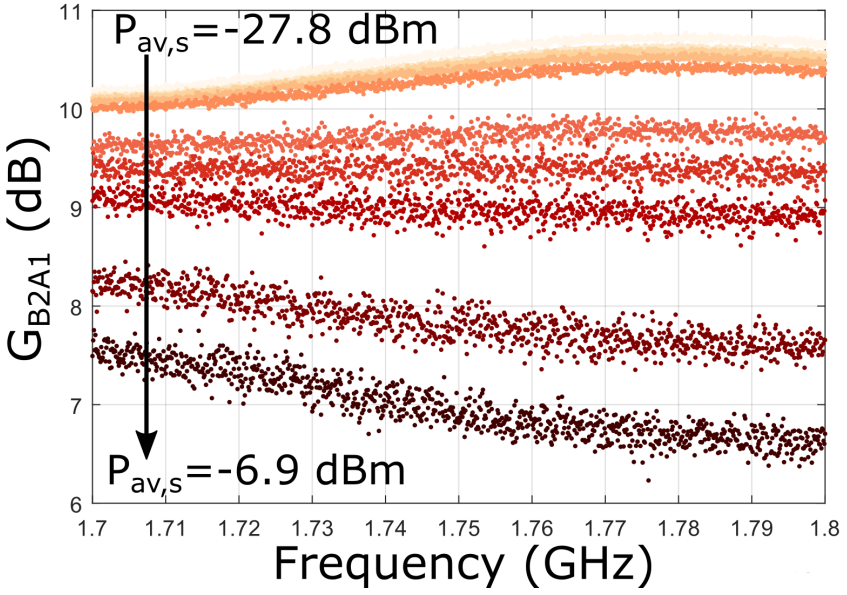


Figure 4.17: $G_{B2A1}(f)$ at different $P_{av,s}$ for the $\Gamma'_{L,2}(f)$ load.

Since $G_{B2A1}(f)$ is identified by averaging out the nonlinear stochastic distortion and that M is fixed for all cases, $G_{B2A1}(f)$ traces show larger spread as $P_{av,s}$ increases: larger values of M would asymptotically result into a continuous $G_{B2A1}(f)$ [113]. The $\mathrm{EVM}_{B_2A_1}$ can be accurately measured below 1%, and its trend was characterized up to 10% for the

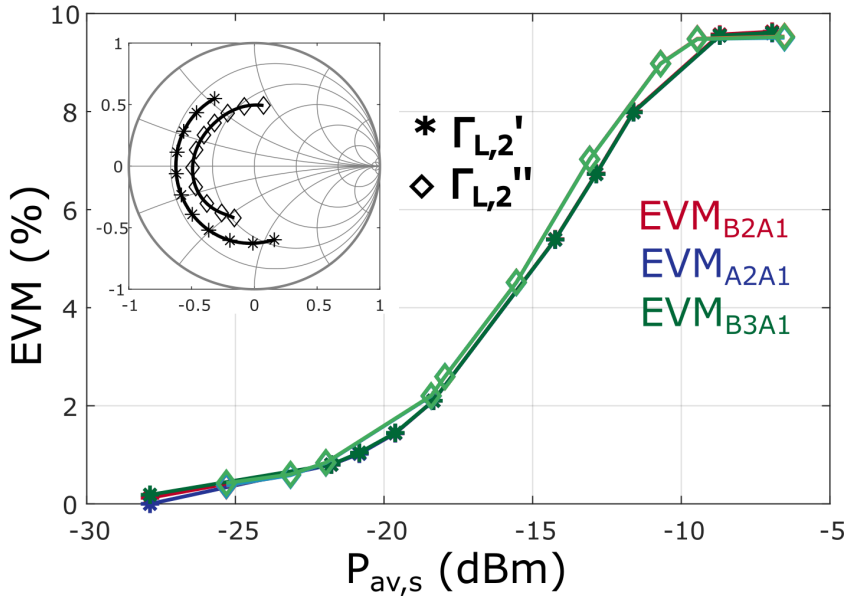


Figure 4.18: EVM profile as a function of $P_{av,s}$ for B_2, A_2 and B_3, for the flat-amplitude loads $\Gamma'_{L,2}(f)$ and $\Gamma''_{L,2}(f)$.

maximum $P_{av,s}$ applied, as shown in Fig. 4.18.
Then, following Sec. 4.3.3 , the procedure has been applied to the case of PISPO-LTI cascade, obtaining the two following EVM characterizations, both using $A_1(f)$ as input:

(A) $\text{EVM}_{A_2A_1}$, where the output $W(f)$ corresponds to $A_2(f)$ and $K(f) = \Gamma_{L,2}(f)$;

(B) $\text{EVM}_{B_3A_1}$ where the output $W(f)$ corresponds to $B_3(f)$ and $K(f) = t_{32}^{21}(f)\Gamma_{L,2}(f) + t_{32}^{22}(f)$.

The results for two load profiles $\Gamma'_{L,2}(f)$ and $\Gamma''_{L,2}(f)$, both nearly flat in amplitude yet with frequency-dependent phase response, as displayed by the tuner, are reported in Fig. 4.18. The three EVM values for a given $\Gamma_{L,2}(f)$ correspond (up to measurement noise), proving that the choice of measurement variable does not impact the actual EVM value for a flat-amplitude load termination. Any of the $B_2(f), A_2(f)$, or $B_3(f)$ can be theoretically used for the characterization, although it is reasonable to choose the one for which the highest acquisition dynamic range can be guaranteed.
Conversely, BLAs differ depending on the output signal, as shown in Fig. 4.19. In fact, in each case, the BLA accounts for the complex transfer function of the cascaded linear network, so that the measurement model of Fig. 4.14b actually corresponds to the general model of a nonlinear-dynamic system (Fig. 4.14a) with the BLA and the stochastic distortion contribution multiplied by $K(f)$, resulting in Fig. 4.14c.

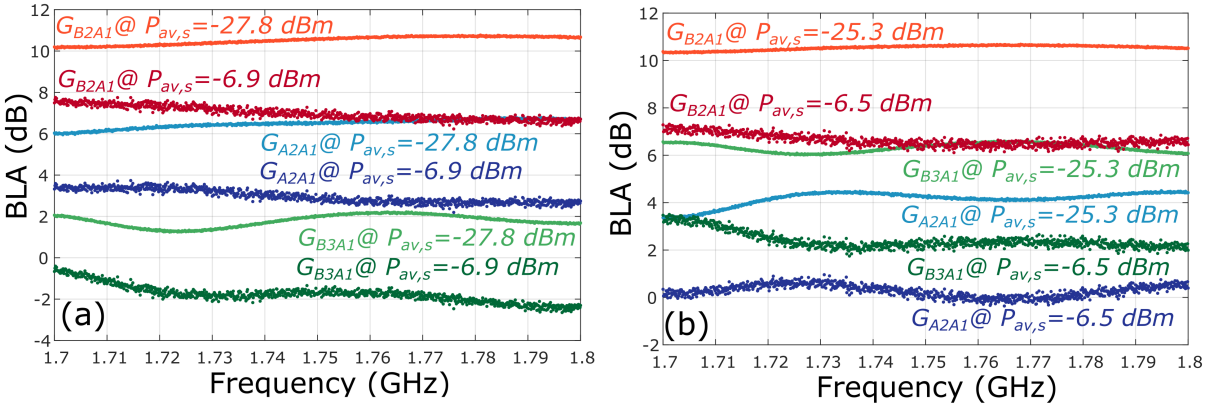


Figure 4.19: $G_{B2A1}(f)$, $G_{A2A1}(f)$ and $G_{B3A1}(f)$ at two different $P_{av,s}$ for a) $\Gamma'_{L,2}(f)$ and b) $\Gamma''_{L,2}(f)$.

When the load is changed to the different flat-amplitude profile $\Gamma''_{L,2}(f)$, also shown in the inset of Fig. 4.18, a different EVM value and corresponding BLAs are measured, as the nonlinear PA will operate at a different large-signal operating point. Nevertheless, also in this second case, the choice of the output variable does not influence the EVM measurement.

4.4 Conclusions

In this chapter, a characterization framework enabling WALP capabilities using frequency-domain VNA acquisitions has been proposed. The approach exploits common VNA HW, removing the typical instantaneous BW requirement found in classical WALP implementations. This allows to significantly increase the available load pull BW which can extend, in principle, up to the full VNA front-end BW. The proposed secants' method allows to iteratively compute the output injected signal required to set a user-prescribed target load reflection coefficient. Its performance compares favorably with other existing numerical algorithms, especially for the large number of tones required to emulate communication standards. This basic method is enhanced using a compensation procedure that accounts for the LS output match presented by the DUT. This network parameter is estimated using a multitone tickler excitation on the output, in a way that is traceable to similar approaches reported in literature. The LS output match compensation is then used to provide a well-conditioned initialization to the secants' iteration. A practical VNA-WALP measurement setup is presented, and characterization results on a packaged GaN HEMT at 5.86 GHz are reported. These results show that the output match compensation step significantly improves stability and convergence speed of the basic secants' algorithm, while handling, at the same time, passive pre-match conditions. The accuracy of the

method is evaluated, highlighting the need for high-dynamic-range VNA acquisitions, and suggesting further improvements in the design of tailored stopping criteria for the iteration. Finally, the application of WALP to out-of-band IM components is discussed and experimentally demonstrated. In principle, the proposed hybrid method including output match compensation can be used to independently synthesize the multitone signals to be injected at different harmonics. Nevertheless, further work and additional HW capabilities are required in order to extend the proposed approach to baseband and harmonic frequencies.

The chapter also reports the performance of a VNA-based, EVM measurement technique in the characterization of modulation distortion of an RF PA under load mismatch conditions. Theoretical derivations and experimental evidence prove that the measured EVM value, while generally depending on the load, is the same for all the waves variables of a flat-amplitude, linear two-port network cascaded with the PA. Differently from other distortion metrics such as ACPR and Noise-Power Ratio NPR, the implemented VNA-based technique clarifies that EVM should be regarded as a network metric of the nonlinear DUT, and not of the signal(s) it is measured on. Future work will involve the VNA-based EVM characterization in the presence of user-imposed broadband load mismatches with frequency-dependent amplitude responses [187].

Chapter 5

Efficient implementation of Volterra behavioral models for radio frequency power amplifiers

5.1 Introduction

Behavioral modeling of radio-frequency (RF) and microwave power amplifiers (PAs) for communications has been the focus of considerable research efforts over the last years [198]–[204]. Multiple physical and circuital causes concur in determining the complex non-linear dynamic behavior displayed by RF PAs under typical operating conditions, e.g., input/output/interstage matching networks, parasitic supply modulation, charge trapping phenomena, and self-heating effects, leading to a variety of interlinked memory effects with characteristic time-constants distributed across many orders of magnitude. Nevertheless, accurate behavioral models of RF PAs are required for system-level simulation and optimization of RF transmitters [106], as well as to improve the design of digital predistortion linearizers [201],[202]. Such models, apart from the good fitting properties, should display favorable characteristics in terms of generality, ease of extraction, and efficiency of the implementation.

Several alternative models to account for non-linear dynamic behavior of RF PAs have been proposed [99]. Many of them (such as memory polynomial and its generalizations [205]) can be seen as tailored truncations of the general Volterra series introduced in Sec. 1.4. With respect to the full series, they typically feature simpler and more direct identification procedures and have formulations tailored to accurately model particular memory effects that might be significant in the PA of interest. Their structure can typically be presented in terms of some combination (series, parallel, feedforward or feedback) of non-linear static and linear dynamic blocks, such as in Wiener-Hammerstein configurations [206].

The modified-Volterra series [207] approaches the problem by expressing the response of a nonlinear dynamical RF PA as a sum of a nonlinear static function (i.e. the AM/AM AM/PM characteristic of the PA) and a series of input-dependent dynamic deviations. Under some relevant assumptions and given the capability to comprehensively account for static nonlinearity and part of the dynamics, this formulation can be, in most practical cases, effectively truncated to the first order [207]. This simplification makes the modified-Volterra approach an attractive choice for system-level simulation, whereas alternative models, often employed with higher-order expressions for DPD [202], are less suitable in this sense due to implementation complexity and unfavorable truncation properties. In this perspective, adopting an high model order will typically imply an ill-conditioned identification, resulting in a signal-dependent behavior that, in turn, is not suitable for system-level simulations.
The first-order modified Volterra model involves the use of nonlinear single-fold integral operators in order to describe the deviation from quasi-static behavior. The effect of such operators is described in terms of time-domain convolutions, or frequency-domain products, of the input with one or more nonlinear (i.e., input dependent) kernels. The same model structure was applied to system-level behavioral modeling of memory effects in RF PAs in [96], and it has been recently extended to the relevant cases of wideband supply [98] and load modulation [86]. More in general, it should be noted that many other similar modeling approaches have been proposed in order to efficiently describe nonlinear memory effects arising in RF PAs [104], [208], [209]. However, independently on their actual structure and represented effects, practically all the models adopt nonlinear integral operators similar to the ones used in the modified-Volterra series. Therefore, the efficient implementation of these operators is paramount for an effective PA simulation at system-level.
Nevertheless, the implementation of these integral operators, either for simulation or linearization purposes, is far from straightforward. Direct time-discretization of the convolution suffers from several shortcomings [208], requiring in general a large number of coefficients and leading to inefficient model implementations. An alternative approach relies on the existence of some particular type of fitting, or decomposition, for the frequency-domain expressions of the nonlinear convolution kernels. This splits the integral operator into a combination of separate static nonlinearities and linear dynamic functions, each of which can be implemented in an efficient way.
Several different strategies of implementation of nonlinear dynamic models for RF PAs have been proposed. For example, in [210] nonlinear dynamic kernels are directly fitted with a set of input-dependent poles and zeros using Padé approximants, leading to a nonlinear infinite impulse response (IIR) implementation of the model. Instead, in [208], [211], a separation of nonlinear dynamic kernels into a sum of products of static nonlinearities and linear dynamics is achieved, using the well-known singular value decomposition (SVD) algorithm. The final implementation is equivalent to a parallel Hammerstein structure (i.e., non-linearity cascaded with a filter).

Instead, the two publications [212], [213], on which the present chapter is based, explore the possibility of having an efficient implementation of the first-order modified-Volterra model for RF PAs by using a dedicated Vector Fitting (VF) algorithm for its nonlinear convolution kernels. VF is a robust numerical method to approximate frequency-domain responses of linear time-invariant systems using poles and residues [214]. Given a frequency response, VF computes a set of stable poles that are real or come in complex conjugate pairs. The versatility of this approach and the widespread use of LTI models led to the application of the algorithm in many different engineering fields, including microwave systems [215]–[218].
The chapter is organized as follows. In Sec. 5.2, the formulation of the first-order modified-Volterra model is reported, analyzing the mathematical and physical constraints impacting its implementation, as well as its empirical identification experiment. Section 5.3 motivates and develops the custom VF procedure and the resulting efficient implementation of the modified-Volterra kernels. In Sec. 5.4, a GaN PA is used to evaluate the fitting performance, analyzing the significance of the various fitting terms in describing the PA behavior. To this aim, other methods [208], [210], [219] are compared on the same dataset. Finally, conclusions are drawn in Sec. 5.5.

5.2 Modified-Volterra PA modeling

5.2.1 Model Formulation

Let $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ be respectively the input and the output passband signals to a RF PA (Fig. 5.1). Their complex-baseband envelopes are respectively named as $x(t)$ and $y(t)$, with the following definitions:

$$\mathcal{X}(t) \stackrel{def}{=} \Re\{x(t)e^{j2\pi f_0 t}\}; \quad \mathcal{Y}(t) \stackrel{def}{=} \Re\{y(t)e^{j2\pi f_0 t}\} \tag{5.1}$$

where f_0 is the carrier frequency in Hz. If the input-output functional that describes the PA behavior can be cast as first-order truncation to its modified-Volterra series [96], the following formulation in terms of complex-envelopes can be adopted:

$$\begin{aligned} y(t) = F[|x(t)|]x(t) + \int_0^{T_m} \tilde{g_1}[|x(t)|,\tau]\{x(t-\tau) - x(t)\}d\tau + \\ + x(t)^2 \int_0^{T_m} \tilde{g_2}'^*[|x(t)|,\tau]\{x(t-\tau)^* - x(t)^*\}d\tau \end{aligned} \tag{5.2}$$

where T_m represents the finite memory duration of the effects in the PA. All the operators in (5.2) automatically satisfy causality and time-invariance constraints. In this form, the model consists of a nonlinear static term, described by the function F which represents the regular AM/AM-AM/PM characteristic of the PA, and two purely dynamic nonlinear

memory terms, described by the nonlinear convolution kernels $\tilde{g}_1$ and $\tilde{g}_2{}'$. This type of first-order truncation of the full modified-Volterra series is valid in two distinct situations, which cover majority of the applications [207]:

- When dynamic deviations $x(t-\tau)-x(t)$ are small within the memory length T_m. This means that memory has to be short with respect to the inverse of the bandwidth of the signals.
- When the purely dynamic memory behavior of the system is mildly nonlinear with respect to $|x(t)|$, and therefore just first order dynamic kernels are sufficient to represent the full PA behavior.

By rearranging (5.2), the following expression is obtained:

$$\begin{aligned} y(t) = F[|x(t)|]x(t) + \int_0^{T_m} \tilde{g}_1[|x(t)|,\tau]\{x(t-\tau)-x(t)\}d\tau + \\ + \frac{x(t)}{x(t)^*}\left[\int_0^{T_m} \tilde{g}_2[|x(t)|,\tau]\{x(t-\tau)-x(t)\}d\tau\right]^*, \end{aligned} \tag{5.3}$$

so that the two kernels $\tilde{g}_1$ and $\tilde{g}_2$ have the same measurement units and the corresponding terms in (5.2) contain identical integrals of the form:

$$\Delta_i(t) \stackrel{def}{=} \int_0^{T_m} \tilde{g}_i[|x(t)|,\tau]\{x(t-\tau)-x(t)\}d\tau \qquad i=1,2. \tag{5.4}$$

Therefore, the system output can be compactly rewritten as:

$$y(t) = F[|x(t)|]x(t) + \Delta_1(t) + e^{j2\arg\{x(t)\}}\Delta_2(t)^*. \tag{5.5}$$

Each $\Delta_i(t)$ in (5.5) can be recast as a corresponding frequency-domain expression:

$$\Delta_i(t) = \int_{-\frac{\mathrm{BW}}{2}}^{+\frac{\mathrm{BW}}{2}} X(f)G_i[|x(t)|,f]e^{j2\pi ft}df \tag{5.6}$$

where:

$$G_i[|x(t)|,f] \stackrel{def}{=} \tilde{G}_i[|x(t)|,f] - \tilde{G}_i[|x(t)|,0]; \quad \tilde{G}_i[|x(t)|,f] \stackrel{def}{=} \mathfrak{F}\{\tilde{g}_i[|x(t)|,\tau]\}(f); \tag{5.7}$$

and BW represents the bandwidth in which the memory effects are considered. Each of the two kernels $G_i[|x(t)|,f]$ has the following properties:

1. It is high-pass ($G_i[|x(t)|,f]_{f=0}=0$). This is coherent with the fact that each $\Delta_i(t)$ describes a purely dynamic deviation that is uniformly zero when the input envelope $x(t)$ is constant.

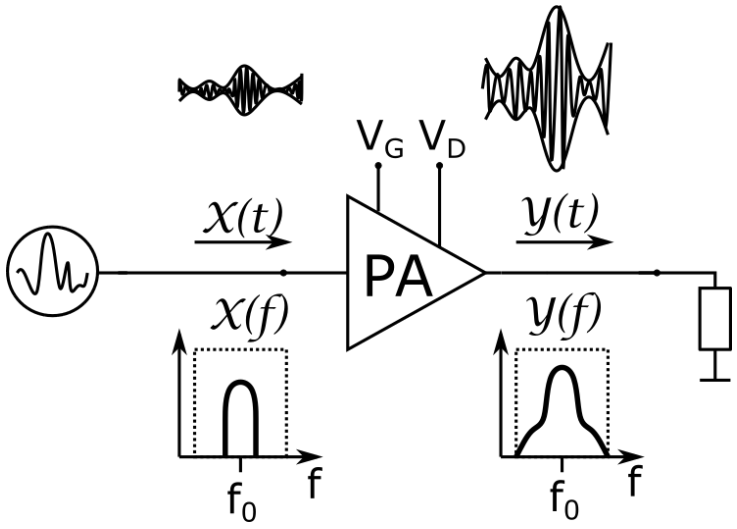


Figure 5.1: Power amplifier behavioral representation in terms of input and output envelopes.

2. It is related to the Fourier transform of a causal complex-valued function of time (i.e., the nonlinear impulse-response g_i). As a consequence, each kernel does not display conjugate-symmetry in the frequency domain ($G_i[|x(t)|, f] \neq G_i[|x(t)|, -f]^*$).

Any suitable numerical implementation of the model will have to comply with these two constraints.

5.2.2 Empirical model identification

The identification of the three functions F, G_1 and G_2 can be carried out in the frequency domain by measuring the PA response to two-tone [96] or three-tone [220] excitations. In the latter case, the input stimulus takes the following form:

$$x(t) = X_0 + \delta_x^+ e^{j2\pi\nu t} + \delta_x^- e^{-j2\pi\nu t} \tag{5.8}$$

where the two "tickle" tones δ_x^+ and δ_x^- at offset frequency $\pm\nu$ are much smaller in amplitude with respect to X_0 setting the large-signal operating point. The PA response is well-approximated by three tones as

$$y(t) \approx Y_0 + \delta_y^+ e^{j2\pi\nu t} + \delta_y^- e^{-j2\pi\nu t} \tag{5.9}$$

where:

$$\begin{aligned}
Y_0 &\stackrel{def}{=} F[|X_0|]X_0; \\
\delta_y^+ &\stackrel{def}{=} \left(F[|X_0|] + \frac{\partial F}{\partial |X_0|}[|X_0|]\frac{|X_0|}{2} + G_1[|X_0|,\nu]\right)\delta_x^+ \\
&+e^{j2arg\{X_0\}}\left(\frac{\partial F}{\partial |X_0|}[|X_0|]\frac{|X_0|}{2} + G_2[|X_0|,-\nu]^*\right)\delta_x^{-*}; \\
\delta_y^- &\stackrel{def}{=} \left(F[|X_0|] + \frac{\partial F}{\partial |X_0|}[|X_0|]\frac{|X_0|}{2} + G_1[|X_0|,-\nu]\right)\delta_x^- + \\
&+e^{j2arg\{X_0\}}\left(\frac{\partial F}{\partial |X_0|}[|X_0|]\frac{|X_0|}{2} + G_2[|X_0|,\nu]^*\right)\delta_x^{+*}.
\end{aligned} \tag{5.10}$$

From the measurements of the three output tones, one finds the values of both kernels, for a given $|X_0|$, ν pair [220]. Then, the large tone amplitude X_0 needs to be swept across the PA operating region of interest, while at the same time the frequency modulation ν ranges within the input signal bandwidth. The overall result of this double sweep is the value of the nonlinear dynamic kernels $G_i[X_0,\nu]$ sampled for a discrete set of points in amplitude $X_0 \in \{X_1,..X_K\}$ and frequency $\nu \in f_1,..,f_R$, where the sampled frequencies can both be positive and negative.

5.2.3 Identification procedure

As demonstrated in the literature, the experimental identification of the modified Volterra model can be performed by means of different measurement setups based on vector network analyzers [220], large-signal vector analyzers [98], nonlinear vector network analyzers [86], and vector signal analyzers [221]. The main advantage of using the three-tone procedure introduced in the previous subsection is that relative measurements with a power calibrated VNA are sufficient to identify all the kernels, without the need for any extra phase information. Indeed, we can recast the the terms in (5.10) as:

$$\begin{aligned}
Y_0 &= F[|X_0|]X_0 \\
\delta_y^+ &= A(|X_0|,\nu)\delta_x^+ + B(|X_0|,\nu)\delta_x^{-*} \\
\delta_y^{-*} &= C(|X_0|,\nu)\delta_x^- + D(|X_0|,\nu)\delta_x^{+*}
\end{aligned} \tag{5.11}$$

In that case, three same-frequency (i.e. LO-invariant) complex-valued ratioed measurements are obtained for each amplitude and frequency point:

$$\begin{aligned}
m_0(|X_0|,\nu) &\stackrel{def}{=} \frac{Y_0}{X_0} = F[|X_0|] \\
m^+(|X_0|,\nu) &\stackrel{def}{=} \frac{\delta_y^+}{\delta_x^+} \\
m^-(|X_0|,\nu) &\stackrel{def}{=} \frac{\delta_y^-}{\delta_x^-}
\end{aligned} \tag{5.12}$$

The measurement m_0 is directly the static characteristic of the PA. Instead, the two measurements at $\pm\nu$ become:

$$\begin{aligned} m^+(|X_0|,\nu) &= A(|X_0|,\nu) + B(|X_0|,\nu)\frac{1}{z^*} \\ m^-(|X_0|,\nu) &= C(|X_0|,\nu) + D(|X_0|,\nu)z \end{aligned} \tag{5.13}$$

where the excitation dependent term z has been defined as $z \stackrel{def}{=} e^{j2arg\{X_0\}}\frac{\delta_x^{+*}}{\delta_x^-}$.
Equation (5.13) is an underdetermined system of two equations in the four unknowns A, B, C and D, which are terms that depend exclusively on the nonlinear dynamic kernels G, as shown by (5.10). In particular, the A and C terms will contain exclusively the direct-kernel values $G_1[|X_0|,\nu]$ and $G_1[|X_0|,-\nu]$ respectively, while B and D contain the values of $G_2[|X_0|,-\nu]^*$ and $G_2[|X_0|,\nu]^*$. Instead, all the information regarding the excitation is grouped in the factor z.
In order to completely determine the four unknowns, two or more sets of measurements (indexed as (1) and (2)) are performed with the same $|X_0|$ and ν. From one measurement to the other, two different sets of small-tones amplitudes and phases are employed, while still keeping the condition $|\delta_x^+|, |\delta_x^-| \ll |X_0|$. Therefore:

$$\begin{aligned} m_1^+(|X_0|,\nu) &= A(|X_0|,\nu) + B(|X_0|,\nu)\frac{1}{z_1^*} \\ m_2^+(|X_0|,\nu) &= A(|X_0|,\nu) + B(|X_0|,\nu)\frac{1}{z_2^*} \\ m_1^-(|X_0|,\nu) &= C(|X_0|,\nu) + D(|X_0|,\nu)z_1 \\ m_2^-(|X_0|,\nu) &= C(|X_0|,\nu) + D(|X_0|,\nu)z_2 \end{aligned} \tag{5.14}$$

The two values of z_1 and z_2 are determined by the known excitation, thus the system is completely solvable to obtain the values $G_1[(|X_0|,\pm\nu)]$ and $G_2[(|X_0|,\pm\nu)]$. The two excitations must be chosen in order to maximize the condition number of the system [220] and reduce the impact of measurement noise. In case that more than two different measurements are performed using different "tickle" amplitudes and phases, the system needs to be solved in the least-square sense.

Impact of non-ideal source

While the three-tone procedure allows to use a regular VNA instead of a fully amplitude and phase calibrated NVNA, non-idealities in the signal source can compromise the identification accuracy.
If the ideal signal exciting the PA is:

$$x_{ideal}(t) = X_0 + \delta_x^+ e^{j2\pi\nu t} + \delta_x^- e^{-j2\pi\nu t} \tag{5.15}$$

the actual one can be represented as:

$$x_{real}(t) = X_0 H(0) + \delta_x^+ H(\nu) e^{j2\pi\nu t} + \delta_x^- H(-\nu) e^{-j2\pi\nu t} \tag{5.16}$$

where $H(f)$ represents a low-pass equivalent linear transfer function of the path linking the "ideal" input signal to the actual one. The excitation-dependent term z has to be substituted with

$$\begin{aligned} z_{real} &= e^{j2arg\{A_0\}} e^{j2arg\{H(0)\}} \frac{\delta_x^{+*} H(\nu)^*}{\delta_x^- H(-\nu)} = \\ &= z \frac{|H(\nu)|}{|H(-\nu)|} e^{j(2arg\{H(0)\} - arg\{H(\nu)\} - arg\{H(-\nu)\})} = z\xi \end{aligned} \tag{5.17}$$

The value of z_{real} differs from the ideal one z because of two factors due to imperfect source behavior: the amplitude imbalance between the lower and upper side band ($|\xi|$) and the in-band deviation from linear phase ($\arg\{\xi\}$). While it is possible to account for the first effect with the amplitude calibration of the VNA, the second effect cannot be easily estimated without the phase information from a wideband reciver. Therefore, even in this VNA-oriented method, the excitation source must display linear phase across all the measurement bandwidth in order to give meaningful results.
For this general case of non-ideal source actuation, the system (5.14) becomes

$$\begin{aligned} m_1^+ &= A(|X_0|, \nu) + B(|X_0|, \nu) = \frac{1}{\xi^* z_1^*} \\ m_2^+ &= A(|X_0|, \nu) + B(|X_0|, \nu) = \frac{1}{\xi^* z_2^*} \\ m_1^- &= C(|X_0|, \nu) + D(|X_0|, \nu)\xi z_1 \\ m_2^- &= C(|X_0|, \nu) + D(|X_0|, \nu)\xi z_2 \end{aligned} \tag{5.18}$$

whose four unknowns are then A, $\frac{B}{\xi^*}$, C and $D\xi$. Even in presence of linear distortion in the signal generator, it is still possible to correctly compute $G_1[|X_0|, \nu]$, while the estimate of $G_2[|X_0|, \nu]$ will inevitably be corrupted by the unknown factor ξ.

5.3 Custom vector fitting implementation

The two dynamic deviations $\Delta_i(t)$ in (5.6) can be evaluated using either time or frequency domain expressions. However, both formulations are poorly suited to a direct implementation:

- Equation (5.4) can be approximated by a non-linear FIR structure [222], [223], by discretizing the time-delay τ and limiting the number of employed coefficients. It is however a rough approximation unless a great number of coefficients is employed.

- Equation (5.6) uses the knowledge of the full spectrum $X(f)$ of the input signal in order to calculate each dynamic deviation at a given time-instant t. In this respect, it is a non-causal implementation of the model, which is not suitable for real-time or CAD simulation applications.

Instead, this work proposes an efficient implementation based on a custom fitting of the nonlinear kernels G_i.

5.3.1 Vector Fitting (VF)

The VF algorithm [214] was conceived to find an optimal partial fraction decomposition of a generic transfer function $M(s)$, describing a stable linear time-invariant (LTI) system in the Laplace domain s, whose frequency response is sampled at discrete positive frequency points f_1,f_2,... f_R. The algorithm efficiently estimates an approximation of the form:

$$M^{VF}(s) = \sum_{k=1}^{Np} \frac{R_k}{s - p_k} \tag{5.19}$$

over the complex s values $j2\pi f_1, j2\pi f_2, \ldots, j2\pi f_R$, by iteratively relocating the N_p poles p_k in (5.19) in order to minimize the weighted error

$$E = \sum_{r=1}^{R} w(f_r)|M^{VF}(j2\pi f_r) - M(j2\pi f_r)|^2 \tag{5.20}$$

where $w(f_r)$ is a strictly positive function that allows to weigh differently errors in different portion of the frequency spectrum. The result of the procedure are the residues R_k and the stable (i.e., negative real part) poles p_k for $k = 1, .., Np$. Due to the hypothesis of a real-valued state space model for the LTI system, the poles and residues are either real or complex-conjugate pairs. The number of poles Np is set in advance by the user to ensure sufficient fitting accuracy. However, particular care must be exercised in avoiding overfitting issues (see Sec. 5.4).
The method (with its standard implementation for linear systems freely available as a MATLAB® routine [224]) can be extended to multi-input multi-output state-space models, in which $M(s)$ becomes a vector of transfer functions. The algorithm then finds a global set of poles that is shared by all the components in the vector, while the corresponding residues represent a relative weight of the various components. This feature of the algorithm can be exploited to provide an efficient implementation of nonlinear dynamic models of the form in (5.6). In this sense, each kernel $G_i[|x(t)|, f]$ can be seen, with respect to the f variable, as a frequency response parametrized by the value of the instantaneous input envelope amplitude $|x(t)|$ [212]. Therefore, for each $|x(t)| = X_k$ in the identification set, $G_i[X_k, f|]$ represents, in general, a different frequency response. While

each response can, in principle, be fitted with an entirely different rational function of f [210], the VF algorithm can be used to determine a set of global fixed poles shared by all the responses (i.e., for all the different X_k) in the identification set. As in the linear multi-input multi-output case, the residue associated with each fixed pole will then be different among the various responses and thus can be seen as a function of the input envelope amplitude X_k.

5.3.2 Vector fitting of modified-Volterra kernels

The two kernels $G_i[|x(t)|, f]$ for $i = 1, 2$ have to be suitably modified in order to be used as an input for the standard VF procedure. Each complex-valued time-domain kernel of (5.4) can be split into its real and imaginary part as:

$$\begin{gathered} g_i[|x(t)|, \tau] = p_i[|x(t)|, \tau] + jq_i[|x(t)|, \tau]; \\ p_i[|x(t)|, f] \stackrel{def}{=} \frac{g_i[|x(t)|, \tau] + g_i[|x(t)|, \tau]^*}{2}; \quad q_i[|x(t)|, f] \stackrel{def}{=} \frac{g_i[|x(t)|, \tau] - g_i[|x(t)|, \tau]^*}{2j}. \end{gathered} \tag{5.21}$$

Due to the causality and stability of g_i, p_i and q_i inherit both properties. In the frequency domain, the previous relations become:

$$\begin{aligned} G_i[|x(t)|, f] &= P_i[|x(t)|, f] + jQ_i[|x(t)|, f]; \\ P_i[|x(t)|, f] &\stackrel{def}{=} \frac{G_i[|x(t)|, f] + G_i[|x(t)|, -f]^*}{2}; \\ Q_i[|x(t)|, f] &\stackrel{def}{=} \frac{G_i[|x(t)|, f] - G_i[|x(t)|, -f]^*}{2j}. \end{aligned} \tag{5.22}$$

The two newly defined kernels display conjugate-symmetry, as they are Fourier transforms of real-valued time-domain functions and therefore are compatible with a standard VF procedure. In order to explicitly account for the high-pass property $P_i[|x(t)|, 0] = Q_i[|x(t)|, 0] = 0$, the kernels are factored as:

$$U_i[|x(t)|, f] \stackrel{def}{=} \frac{P_i[|x(t)|, f]}{j2\pi f}; \quad V_i[|x(t)|, f] \stackrel{def}{=} \frac{Q_i[|x(t)|, f]}{j2\pi f} \tag{5.23}$$

where U_i and V_i are not anymore constrained to be high-pass. The VF algorithm can then be applied to U_i and V_i in order to compute a joint partial fraction decomposition for the two functions, using the measured values of the original complex kernel $G_i[X_k, f_l]$ for $X_k \in \{X_1, ..X_K\}$ and $f_l \in \{-f_L, .. - f_1, f_1.., f_L\}$. This results in:

$$U_i^{VF}[|x(t)|, s] = \sum_{k=1}^{Np_i} \frac{u'_{i,k}[|x(t)|]}{s - p_{i,k}}; \quad V_i^{VF}[|x(t)|, s] = \sum_{k=1}^{Np_i} \frac{v'_{i,k}[|x(t)|]}{s - p_{i,k}} \tag{5.24}$$

where

$$U_i^{VF}[|x(t)|,s]_{s=j2\pi f} \approx U_i[|x(t)|,f]; \quad V_i^{VF}[|x(t)|,s]_{s=j2\pi f} \approx V_i[|x(t)|,f] \tag{5.25}$$

This procedure splits the components of the nonlinear dynamic kernels into purely static nonlinear terms, described by the input-dependent residues $u'_{i,k}[|x(t)|]$ and $v'_{i,k}[|x(t)|]$, and purely linear dynamic terms, corresponding to the globally shared set of poles $p_{i,k}$. The residues, in this respect, work as "activation" functions of the different elementary dynamics, modeled by the poles. The division by f in (5.23) tends to overemphasize the low-frequency memory behavior in most cases. The typical result is the presence of one or more poles located at frequencies much lower than the ones at which the model has been identified. This issue can be solved by proper use of the weighing function, which can be designed in order to correct this behavior while at the same time retaining a strict high-pass property of the kernels.
All the obtained stable poles are either real or in complex-conjugate pairs, while the corresponding residues are respectively real or complex-conjugate. This allows to rearrange the terms in (5.24) to yield a set of Np_i physically realizable (i.e., causal and stable) first and second-order linear filters $H_{i,k}(s)$ and two sets of real-valued residues $u_{i,k}[|x(t)|]$ and $v_{i,k}[|x(t)|]$.

$$U_i^{VF}[|x(t)|,s] = \sum_{k=1}^{Np_i} u_{i,k}[|x(t)|]H_{i,k}(s); \quad V_i^{VF}[|x(t)|,s] = \sum_{k=1}^{Np_i} v_{i,k}[|x(t)|]H_{i,k}(s) \tag{5.26}$$

The use of real and complex-conjugate poles to fit non-conjugate-symmetric kernels is alternative to the approach reported in [218], in which VF is employed to represent complex-baseband equivalents of bandpass linear S-parameters. In that case, non-conjugate-symmetric frequency domain data is handled by modifying the original algorithm to allow for the existence of purely complex poles, without any associate complex-conjugate companion. This choice, however, is poorly suited for the implementation of the model in commercially available CAD tools, which can simulate just physically realizable (i.e., real-valued) systems.
Finally, a fitting function G_i^{VF} for the original high-pass non-conjugate symmetric kernel G_i is found as:

$$\begin{aligned} G_i^{VF}[|x(t)|,s] &= s(U_i^{VF}[|x(t)|,s] + jV_i^{VF}[|x(t)|,s]) = \\ &= \sum_{k=1}^{Np_i} u_{i,k}[|x(t)|]sH_{i,k}(s) + j\sum_{k=1}^{Np_i} v_{i,k}[|x(t)|]sH_{i,k}(s) = \\ &= \sum_{k=1}^{Np_i} u_{i,k}[|x(t)|]L_{i,k}(s) + j\sum_{k=1}^{Np_i} v_{i,k}[|x(t)|]L_{i,k}(s) \end{aligned} \tag{5.27}$$

where $L_{i,k}(s) = sH_{i,k}(s)$ are first and second order high-pass and band-pass linear filters. Each $L_{i,k}(s)$ is both stable and causal because it is represented by a proper transfer function, as each $H_{i,k}(s)$ is strictly proper from (5.26), with poles in the open left-hand plane. The final form reached in (5.27) describes a fitting function that satisfies the required constraints: it is high-pass (0 for $s = j2\pi f = 0$) and causal (the degree of the numerator does not exceed the degree of the denominator).
The separation between nonlinearity and dynamics yields a particularly compact expression for each of the two integrals in (5.2), once the input envelope $x(t)$ is split into its IQ components as $x(t) = x_p(t) + jx_q(t)$:

$$\Delta_i(t) = \int_{-\frac{BW}{2}}^{+\frac{BW}{2}} G_i[|x(t)|, f]X(f)e^{j2\pi ft}df \approx \int_{-\frac{BW}{2}}^{+\frac{BW}{2}} G_i^{VF}[|x(t)|, f]X(f)e^{j2\pi ft}df =$$
$$= \sum_{k=1}^{Np_i}\left[u_{i,k}[|x(t)|]\cdot\int_{-\frac{BW}{2}}^{+\frac{BW}{2}} L_{i,k}(j2\pi f)X_p(f)df - v_{i,k}[|x(t)|]\cdot\int_{-\frac{BW}{2}}^{+\frac{BW}{2}} L_{i,k}(j2\pi f)X_q(f)df\right] +$$
$$+j\sum_{k=1}^{Np_i}\left[u_{i,k}[|x(t)|]\cdot\int_{-\frac{BW}{2}}^{+\frac{BW}{2}} L_{i,k}(j2\pi f)X_q(f)df + v_{i,k}[|x(t)|]\cdot\int_{-\frac{BW}{2}}^{+\frac{BW}{2}} L_{i,k}(j2\pi f)X_p(f)df\right] =$$
$$\stackrel{def}{=} \sum_{k=1}^{Np_i}\delta p_{i,k} + j\sum_{k=1}^{Np_i}\delta q_{i,k} \tag{5.28}$$

5.3.3 Model implementation details

The implementation of each Δ_i involves the sum of Np_i parallel branches. In each branch the input envelope $x(t)$ is separately processed by the linear filter $L_{i,k}(j2\pi f)$ and by the nonlinear residues $u_{i,k}[|x(t)|]$ and $v_{i,k}[x(t)]$. The results of these two operations are mixed together and combined to yield the overall dynamic deviation. This structure is easily implementable in CAD or system simulators as the block diagram in Fig. 5.2, using, for each branch:

- A single first or second order linear filter (e.g., RLC networks).
- Two look-up tables (LUTs) for the nonlinear residues functions.
- Four ideal mixers.
- Two Adders/Subtracters.

The overall feedforward model has been shown to be compatible with harmonic balance and envelope transient simulations [212]. Computation of the linear filter responses can be

sped up by their implementation in the discrete-time domain as finite-impulse-response stable filters [208], [223].
A more compact representation can be achieved by using a common set of N_p poles p_k (where the index i is dropped) for both G_1 and G_2. This is a reasonable simplification allowed by physical insight: memory effects with time constants within the modulation bandwidth will likely affect the values of both kernels (see Section 5.4). In this case, the overall flow of the fitting process can be summed up in the following steps:

1. Identification measurements for the extraction of the bivariate nonlinear dynamic kernels G_1 and G_2 across the required PA operating range (large-signal operating point and bandwidth).

2. Computation of auxiliary kernels P_i, Q_i, U_i, V_i that preserve the salient characteristics of the original ones and that can be vector-fitted using standard algorithms.

3. Computation of fixed global poles and input-dependent residues using VF.

4. Computation of the LUTs and elementary filters.

5.4 Performance evaluation

To demonstrate the implementation of Sec. 5.3, we apply the modified-Volterra model identification procedure to a nonlinear RF PA by means of simulation-based experiments. In particular, the device-under-test (DUT) is a one-stage PA based on a 6-W GaN-HEMT (Cree CGH40006P), operated at a carrier frequency of 2 GHz, where the simulated equivalent-circuit model of the transistor has been provided by the manufacturer. Just as in [212], the output bias network was intentionally modified with respect to an optimal design in order to enhance memory effects in the band of interest. This configuration corresponds to a realistic situation where the bias network is the main cause of unwanted dynamic effects within the operating bandwidth.
The nonlinear dynamic kernels G_1 and G_2 are extracted using the three-tone identification method depicted in Sec. 5.2. The identification region spans 80 MHz of modulation bandwidth swept with a 400-kHz frequency step, resulting in a memory span $T_m \leq 2.5\ \mu s$, and up to 5 dB of static compression in terms of amplitude. If the first-order truncation could not provide sufficient modeling accuracy for nonlinear dynamic effects involved within T_m, e.g., due to the presence of strongly nonlinear low-frequency (LF) phenomena induced by the resonances of the bias network, a multi-input modified-Volterra approach as in [98] could be adopted, where the supply voltage is treated as a separate input variable. The VF implementation described in Sec. 5.3 would be equivalently applicable to the multi-input case, as the additional LF-to-RF transfer function can be processed in the same way as the RF-to-RF ones.

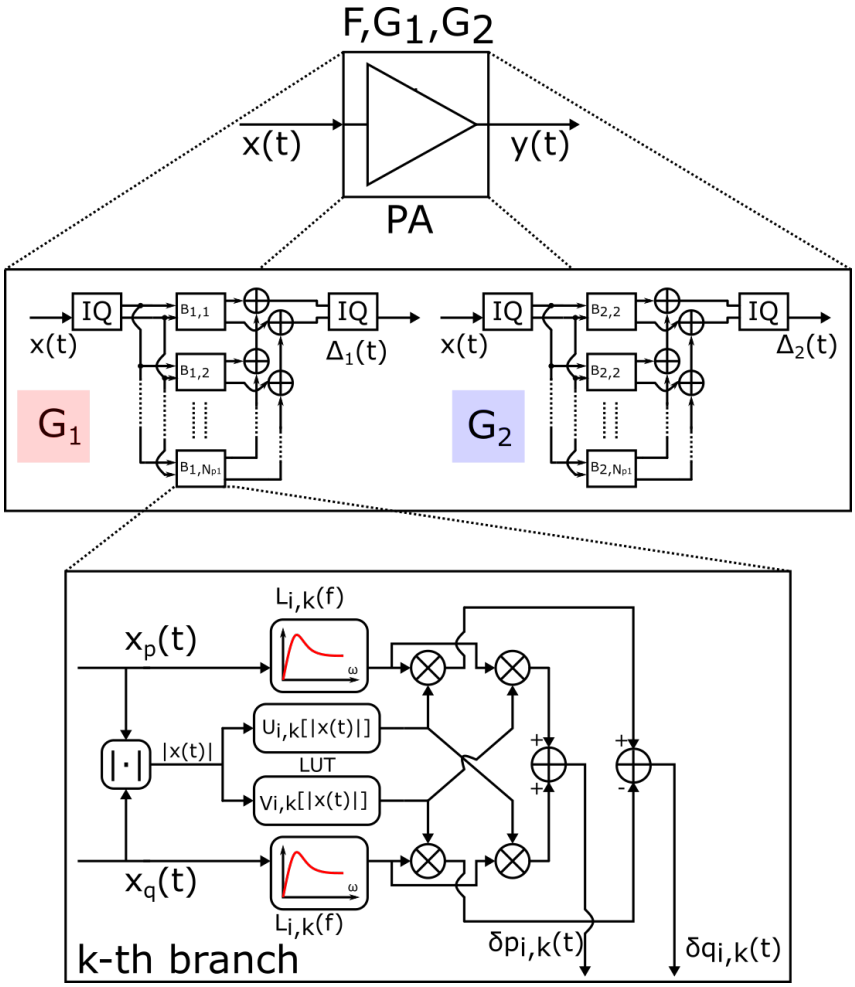


Figure 5.2: Block diagram of the implementation of the k-th branch of the nonlinear dynamic kernel.

The large signal input is swept at closely spaced intervals (>100 steps) to correctly estimate not just the value of $F[|X_0|]$, but also of the derivative $\frac{\partial F[|X_0|]}{\partial |X_0|}$ at each point, which, due to (5.10), is fundamental for the correct de-embedding of purely dynamic kernels from the identification measurements. The accurate kernel identification is critical to avoid vector-fitted models with high variance and poor generalization capability. Considering that the model extraction procedure depicted in Sec. 5.2.2 is defined in the frequency domain, exploiting high-dynamic range instrumentation like network analyzers allows obtaining accurate and low-noise kernel identification. Otherwise, in the presence of noisy kernels data, weighting techniques such as the ones in [217] should be adopted. Then, the objective of the implementation is accurately reproducing these two extracted bivariate kernels, which are shown in Fig. 5.3 on the amplitude and frequency axes.

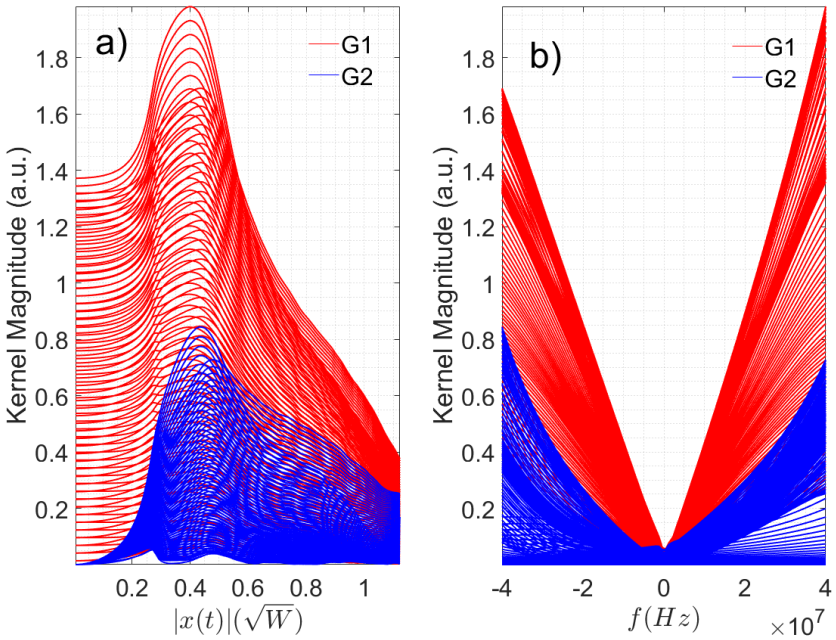


Figure 5.3: Projection of G_1 and G_2 on the amplitude (a) and frequency (b) axes for the simulated PA

The custom VF routine is applied in MATLAB® to both kernels. In order to assess the performance of the fitting and gauge the correct number of poles, the normalized mean square error (NMSE) can be employed as a useful metric, with the following definition:

$$NMSE_i \stackrel{def}{=} \frac{\sum_{k=1}^{K}\sum_{r=1}^{R} |G_i^{VF}[|X_k|, f_r] - G_i[|X_k|, f_r]|^2}{\sum_{k=1}^{K}\sum_{r=1}^{R} |G_i[|X_k|, f_r]|^2}. \tag{5.29}$$

The number of poles is sequentially increased until the NMSE in fitting the kernels is sufficiently reduced, or a maximum model-order complexity is reached. The use of an excessive number of poles (as discussed in Sec. 5.4.1) could result in ill-conditioned residue functions with strong oscillating behavior with respect to input amplitude. While this iterative approach is not suitable for automated model-order selection, the procedure is performed in an offline fashion directly on the available kernels and typically presents a very low computational cost.

The following sections further discuss the proposed model and its features are compared with other implementations of nonlinear-dynamic kernels found in literature. In particular the frequency-domain direct LUT implementation of (5.6), the nonlinear FIR-IIR struc-

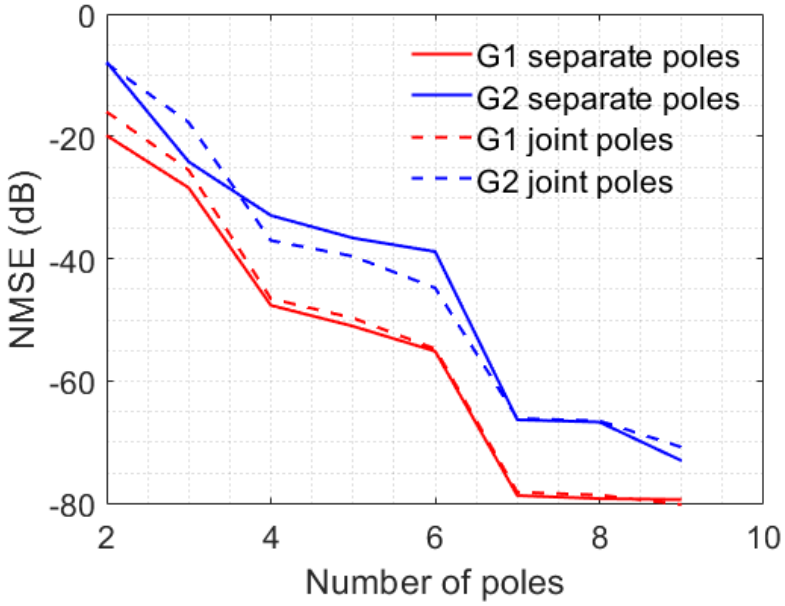


Figure 5.4: NMSE for different numbers of fitted poles in case of separate poles for G_1 and G_2 (solid lines) or a common set of poles (dashed lines).

tures [219][210] and the Singular Value Decomposition (SVD) approaches are considered. As shown in Table 5.1, the performance in terms of number of poles and residues, the use of global poles and the overall implementability of the model is analyzed.

	FD-LUT (5.6)	NL FIR-IIR (5.4) [219][210]	SVD [208][211]	This work
Fixed poles	-	No	Yes	Yes
# Poles	-	-	Large	Small
# NL Functions	-	Large	Small	# Poles
Causality	No	Yes	Yes	Yes
Stability	Yes	?	Yes	Yes

Table 5.1: Comparison with existing implementation methods for nonlinear dynamic kernels

5.4.1 Number and position of poles

The NMSE resulting from a different number of poles is is shown in Fig. 5.4, where two different cases are considered, first using the same set of joint poles for both kernels ($p_{1,k} = p_{2,k}$), then fitting each kernel with a separate set of poles ($p_{1,k} \neq p_{2,k}$). The use of additional poles brings just a slight reduction in the NMSE in some cases (e.g., from

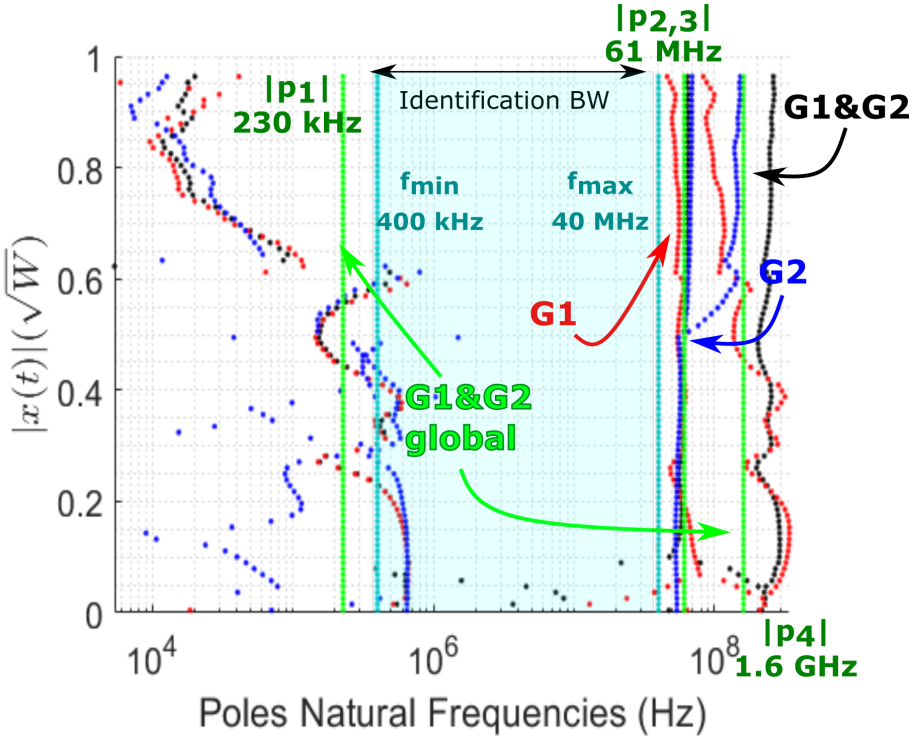


Figure 5.5: Locations of the $Np = 4$ fitted poles at different amplitudes for G_1 (red), G_2 (blue), jointly for G_1 and G_2 (black) and the global poles proposed in this work (green). The characteristic frequencies of the four global poles are also reported (dark green).

4 to 5 poles), while the improvement is significant in others (e.g. from 3 to 4 poles). The number of poles has then to be picked as a trade-off between the number of coefficients and acceptable NMSE. Moreover, the choice of a separate set of poles for each kernel shows little improvement in performance in terms of NMSE, while effectively doubling the set of coefficients with respect to a common set of poles. In this particular case, a number of $N_p = 4$ shared poles ensured a NMSE of almost -40 dB, which is here considered as acceptable for most applications. Figure 5.5 compares different possible choices for the fitting strategy, for a fixed $N_p = 4$:

1. The use of global poles shared by both kernels as proposed in this work and in [212].
2. The use of a different set of input-dependent poles for each kernel, in an approach similar to [219] and [210] underlying to a nonlinear FIR-IIR implementation structure.
3. The use of common set of input-dependent poles shared by both kernels, again as a modification of [219] and [210].

In all the considered cases, the VF algorithm produces two real poles and a complex-conjugate pair. The input-dependent pole at lowest frequency displays an ill-defined behavior, as the impact of its exact location is largely irrelevant on the kernel shape within the considered bandwidth. The other three poles are located above the modulation bandwidth for which the kernels have been identified. This fact is a direct consequence of the kernels' concave shape when projected upon the frequency axis (Fig.5.3b). This behavior is to be expected if the considered PA can be actually modeled with a first-order truncation of a modified-Volterra model. In fact, the fundamental short-memory assumption of the model (see Sec. 5.1) is that dynamical effects have time constants, represented in this context by the reciprocals of the fitted poles frequencies, which are significantly shorter than the reciprocal of the input signal bandwidth. The locations of the three high-frequency input-dependent poles tend to continuously drift away with increasing input amplitude. In this respect, the previous observations suggest that the fitting poles, in all the approaches, have to be interpreted just as a result of an optimization procedure, giving nothing more than rough indications of the frequency ranges in which significant memory behavior is observed. While it might be tempting to seek a correspondence between characteristic frequencies of the fitted poles and physical causes of memory, the complexity of the actual PA behavior defies this type of description.
For the PA under exam, the choice of the VF strategy using fixed-location global poles seems to be reasonable. This approach yields well-defined (i.e., input-independent) approximated estimates of the involved memory time-constants and, as outlined by the previous NMSE evaluation, displays a high-degree of accuracy in fitting the original bivariate kernels. Moreover, the proposed implementation is considerably simpler than the nonlinear IIR filters [210], [219] resulting from the input-dependent poles approach. The elementary filters and corresponding residues resulting from the global poles approximation of each of the two kernels are shown in Fig. 5.6. As previously mentioned, the residues are input-dependent weighing functions for the respective elementary filters. Therefore, their values can be used to estimate the influence of each pole at a given amplitude and eventually "prune" redundant poles, if the corresponding residue is uniformly close to zero. In this case, the residues for the three high frequency poles are relatively smooth function of the input signal amplitude. Instead, the residue associated with the lowest frequency pole shows significant variation, which can however be neglected due the negligible value of the residue itself with respect to the others. In general, the use of an overly large number of unnecessary poles in the fitting tends to produce wildly varying residues, as the optimization algorithm solves a largely over-determined system.

5.4.2 Comparison with singular value decomposition (SVD)

The proposed VF method was compared with the one shown in [208],[211], accordingly adapted for the case in exam. In [211], a non-linear dynamic kernel is decomposed using SVD into a sum of products of static non-linearities $T_k(|x(t)|)$ and linear frequency-

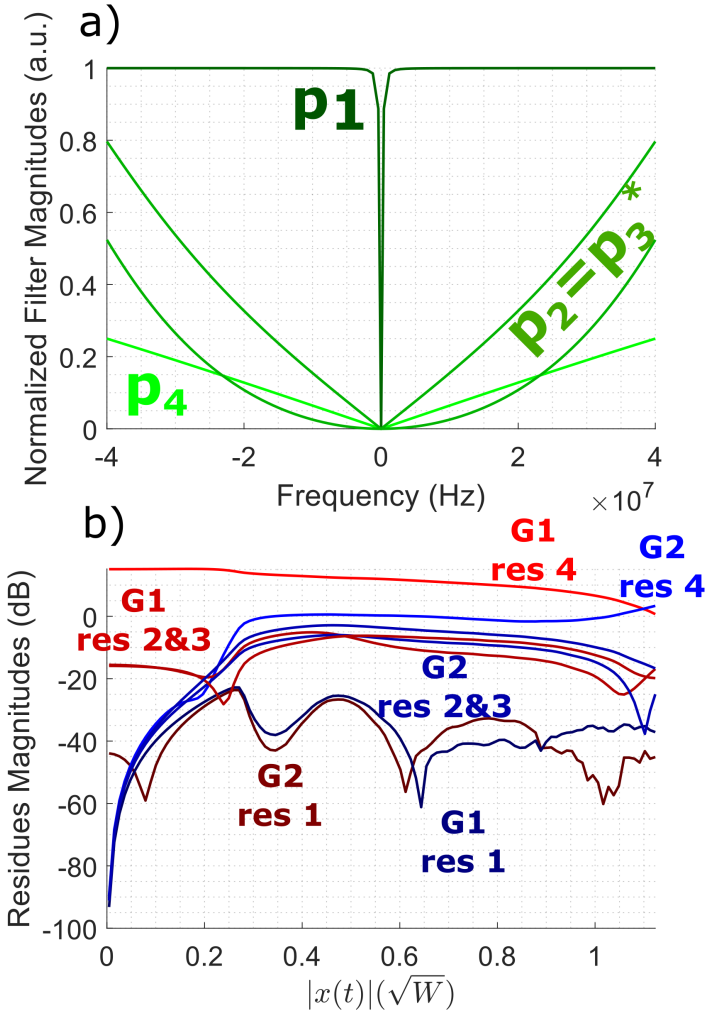


Figure 5.6: (a) Elementary first and second order filters associated with the poles and (b) corresponding residues for the two kernels.

responses $J_k(f)$ of the form:

$$G_i[|x(t), f] = \sum_{k=0}^{N_{SVD}} \sigma_k T_k(|x(t)|) J_k(f), \tag{5.30}$$

where N_{SVD} is picked in order to include in the fitting all the significant singular values σ_k. Each frequency response $J_k(f)$ is fitted using polynomials in the variable f, which maps into an implementation using time derivatives of the input signal. Instead, VF of the frequency responses is shown in [208] using, in the general case, a different set of poles for each k-th function. In this second case, the separation between nonlinear functions

Table 5.2: NMSE of the proposed nonlinear dynamic model implementation and the implementation in .

N_{NL}	N_P	G_1^{VF}	G_2^{VF}	G_1^{SVD}	G_2^{SVD}
2	2	-16	-8	-15	-18
	4			-29	-24
	6			-30	-24
	8			-32	-24
4	4	-47	-37	-13	-17
	6			-23	-29
	8			-33	-38
	10			-46	-45

and linear dynamics closely resembles the one in (5.26). The main difference lies in the fact that each response $J_k(f)$ is described using different set of multiple (typically more than two) poles.
The NMSE resulting from the two approaches, i.e., the one adopted this work and [208], applied to the nonlinear kernels of the PA in exam, is shown in Tab. 5.2 for up to four nonlinear static functions (N_{NL}) and number of poles used to fit linear dynamics (N_P). By construction, N_P must correspond to N_{NL} (residues) in VF, whereas in SVD, N_P can be chosen independently.
The SVD algorithm ensures a minimal number of terms in the sum of (5.30), achieving satisfactory NMSEs using a reduced number of nonlinear functions. However, the resulting filters are constrained to be orthogonal in the frequency domain, requiring in general a large number of poles which cannot be reused among different linear responses.
The proposed VF-based algorithm, instead, assumes that the kernels are efficiently described by residue functions that weigh elementary (i.e., first or second order) frequency responses depending on the input amplitude. This approach has the potential to be particularly efficient when the dynamical behavior of the kernels can be described using very few time-constants, which show little dependence on the PA large signal operating point. However, since the number of input-dependent residues is constrained to be equal to the number of poles used in the fitting, the implementation may result in a potentially greater number of nonlinear functions.

5.4.3 Large signal validation

The objective of the validation consists of verifying that both the frequency-domain, non-causal LUT-based implementation and the proposed efficient implementation produce

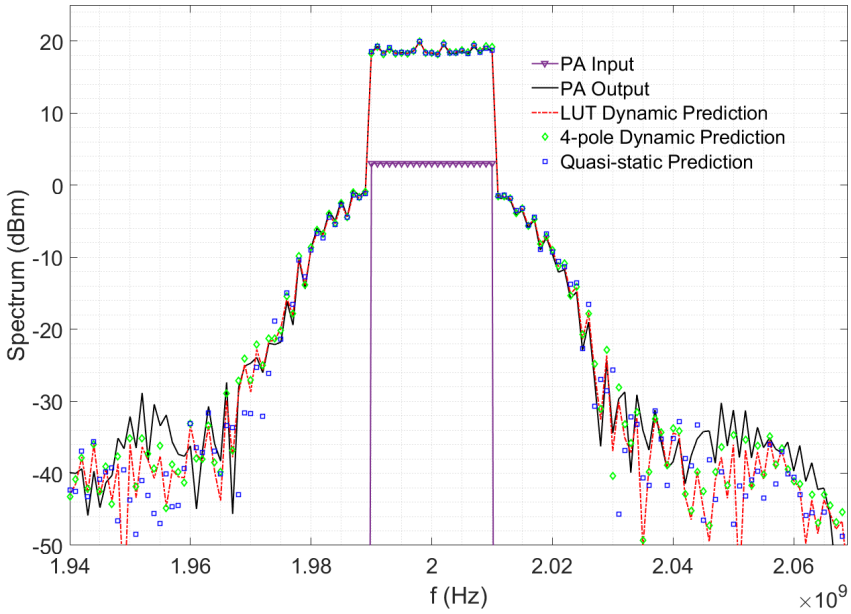


Figure 5.7: Input, output and modeled spectra for the PA under a 20-MHz-wide multitone modulation.

exactly the same results in an application-like large-signal regime.
The proposed model structure was implemented in Keysight ADS® circuit-simulation environment, using standard Symbolic Defined Devices components for the the mixers, adders, nonlinear residual functions and built-in linear filters for the single-pole sections. Both this implementation and the original PA circuit were evaluated through envelope-transient simulations in ADS. Instead, the LUT-based implementation was evaluated in MATLAB®, as the non-causality of the time-frequency operator prevents its use in circuit-level simulations.
The PA was excited with a random-phase 20-tone signal occupying a bandwidth of 20 MHz and displaying a PAPR of 10.4 dB. The power spectra of the input and output signals, together with quasi-static, LUT-based modified-Volterra and four-pole vector-fitted modified Volterra model predictions are reported in Fig. 5.7. The corresponding time-domain envelopes are shown in Fig. 5.8. Finally, the dynamic AM/AM characteristic of the PA is reported in Fig. 5.9. The PA shows significant hysteresis for this level of compression and modulation bandwidth, with the overall behavior departing significantly from the quasi-static model prediction. Nevertheless, both the LUT-based modified-Volterra model and its vector-fitted approximation using four poles are able to accurately predict the PA response with similar accuracy, as expected from the NMSE results.

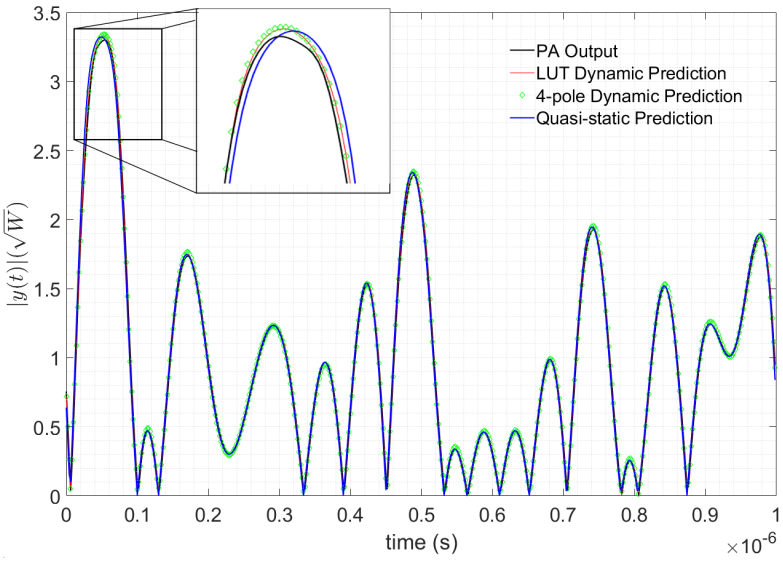


Figure 5.8: Output and modeled time-domain envelopes for the PA under a 20-MHz-wide multitone modulation.

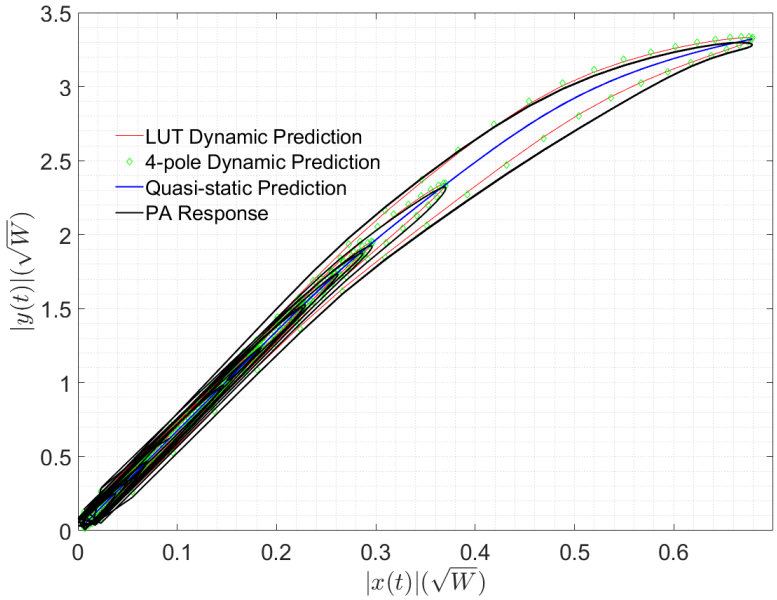


Figure 5.9: Dynamic AM/AM characteristics of the PA under a 20-MHz-wide multitone modulation.

5.5 Conclusions

This chapter describes an efficient system-level implementation technique for single-fold convolution integrals with nonlinear memory kernels, a mathematical construct commonly

adopted in nonlinear dynamic PA behavioral models. In particular, the widely adopted modified Volterra formulation was considered. A custom implementation of the VF algorithm is employed to provide an approximation of the kernels, which results in their decomposition in terms of low-order linear dynamics and simple static nonlinearities. The use of a global set of poles in the fitting allows a causal and stable implementation of this model in CAD or system-level simulation environments using common components. The resulting structure is compatible both with harmonic-balance and envelope-transient simulators, and the reported performance in terms of compactness and ease of implementation compares favourably with other existing methods.
A commercial GaN PA is used to validate the approach, with the proposed method yielding an extremely compact representation which accurately fits the predicted model behavior.

Chapter 6

Conclusions

6.1 Main achievements

The Ph.D. research reported in this manuscript has concentrated on different aspects in the characterization and modeling of GaN HEMTs and power amplifiers for radio-frequency applications, with a focus on the experimental investigation of trapping and memory phenomena. The results highlight the significant impact of low-frequency dynamics on the electrical behavior of GaN devices, particularly in the broadband modulated operating conditions encountered in modern telecommunications standards. The study of these effects has required the development of several novel large-signal measurement setups and related characterization techniques, particularly at RF frequencies.
The research activity produced the following results:

- A low-frequency oscilloscope-based measurement setup for the time-domain characterization of trapping phenomena in GaN HEMTs has been developed. The bench enables the acquisition of long current transients under arbitrary gate and drain excitations, maintaining at the same time the pulsed IV capabilities present in traditional setups. Moreover, a novel time-domain global trapping characterization method, based on 50%-duty cycle large signal pulses, has been proposed. The setup and related techniques were used to characterize charge capture and release dynamics in several short-gate-length GaN devices for microwave and sub-mmwave application. The investigations allowed to highlight different physical phenomena, together with the time-constants and activation energies of interest. The results are published in [54], [130].

- The impact of trapping and other memory effects on the RF operating characteristics of GaN devices and power amplifiers has been investigated using different specific large signal measurement setups and methods. First, the double pulsed characterization technique and bench, originally devised in order to obtain IV characteristics

of GaN HEMTs at a fixed trapping state, has been extended in order to enable the measurement of iso-dynamic S-parameters of the DUT. Then, swept two-tone and pulsed large signal RF measurements have been compared with respect to their ability of highlighting different trapping behaviors in GaN HEMTs. Finally, a novel technique based on the best-linear-approximation framework has been used in order to characterize the effect of trapping directly on PA performance. The method involves the measurement of variations in the broadband modulation distortion performance of the PA when the tone spacing in a standard-compliant multisine excitation is varied in order to probe dynamical effects across different time scales. The results are published in [69], [172], [173].

- A novel wideband active load pull measurement setup based on a vector network analyzer platform has been devised, developed and validated. The proposed frequency-domain method avoids the use of wideband acquisition hardware and the implementation of phase-coherent RF acquisitions. The setup synthesizes the required broadband load by a combination of a modified secants' numerical technique and a measurement-based compensation of the large signal output match of the device under test. Moreover, the impact of load pull conditions on the measurement of broadband modulation distortion metrics has been theoretically and experimentally evaluated. The results are published in [185]–[187].

- An efficient implementation technique for the modified-Volterra behavioral model of RF PAs subject to memory effects has been proposed. Through the use of a modified vector fitting technique, the single-fold nonlinear-dynamic convolution operators are decomposed in a set of static nonlinearities and linear filters, which can be compactly realized in CAD environments. Simulation results on a GaN PA are used in order to validate the method and compare it with existing implementation techniques. The results are published in [212], [213].

6.2 Future work

The results obtained during the Ph.D. still leave many further developments to be investigated, with several open challenges both at device and PA level. The following sections give a brief outline of possible next steps, some of which are already being actively pursued by various research groups in academia and industry.

6.2.1 Empirical modeling of trapping phenomena

The global trap dynamics of modern GaN HEMT technologies has been shown to be poorly characterized by many existing methods and model formulations, particularly under large signal excitations. The variety of observed behaviors points to the necessity of

novel general-purpose descriptions for trapping phenomena in order to improve behavioral and equivalent circuit models in design kits for PA applications.
It would indeed be interesting to study the possibility of developing "grey-box" behavioral trap models [99] directly from an extended base of measurements data from existing techniques, such as pulsed current transients, LF S-parameters or LF load pull. The resulting empirical formulation could describe the observed behavior in terms of the time-domain evolution of equivalent state variables that jointly describe trap and thermal dynamics [225]. This type of model should have the main objective of reproducing the observed trap behavior in a large variety of conditions, possibly using different excitations than the ones adopted for the identification. To this goal, analytical design of experiments approaches [226] can be leveraged in order to select a meaningful dataset to be used for model extraction. At the same time, this type of empirical model should still offer some level of physical interpretability of the effects and maintain a connection with classical descriptions used in semiconductor literature [47], [135].
A similar research line could investigate the implications of fitting this type of trapping behavioral models directly on RF data [59], [227], in order to better represent the final application conditions. While LF models are intrinsically valuable and a fundamental component of many RF ones [228], the description of the LF-to-RF conversion of trapping effects is expected to be extremely challenging.
In this respect, the pulsed/double-pulsed transient small-signal S-parameter measurements proposed in Section 3.2 can be used to assess the impact of trapping on the various components of quasi-static GaN-HEMT models. On top of the clear effect on the intrinsic current generator, this approach can also shed light on the dependence of capacitances/charge-functions [174] on the trapping state, which is comparatively less reported in literature [175]. Thanks to the flexibility of the proposed setups, this characterization approach can be also extended in order to use different low-frequency excitation strategies, such as the multi-bias 50%-duty-cycle pulses introduced in Sec. 2.4, in order to obtain a more global description of trapping.
At the same time, the direct characterization of GaN HEMTs and PAs under large-signal RF-modulated signals is of interest in the development of charge-trapping models, particularly using pulsed-RF [113] and multitone excitations [87], [173]. The extension of the proposed WALP setup to an on-wafer measurement environment has the possibility of enabling the collection of complete datasets for this type of modeling. Indeed, load pull capabilities have been shown to be crucial in reproducing realistic large-signal load-line operating conditions and are expected to provide valuable information on the trapping and thermal state in GaN devices [171], [229].

6.2.2 Wideband active load pull

In the proposed VNA-WALP approach, the secants method has shown sub-optimal performance in terms of convergence speed and stability, particularly in low-SNR scenarios.

Further study is required in order to devise new algorithms and stopping criteria that can improve the overall robustness of the technique.
Thanks to the use of regular VNA hw, the VNA-WALP method is well suited to be combined with the BLA-based EVM measurement framework used throughout the thesis work. In this way, it would be possible to emulate matching conditions encountered in the final applications, allowing to asses the broadband modulated distortion of a given device (e.g. constant-EVM loci on the Smith chart) without having to realize any physical PA circuit. Some preliminary results in this respect are reported for an matched DUT in an upcoming publication [230] (realized after the end of this PhD), analyzing the performance of an improved VNA-WALP algorithm and its integration with a commercial VNA-based EVM measurement platform. This combination between EVM and WALP characterizations also allows to study more in detail the proposed EVM model in actively synthesized mismatch conditions, where non-flat loads can more easily be realized in order to further validate the theoretical predictions of the framework.
From a practical perspective, the implementation of VNA-WALP for higher order IM, harmonic and especially baseband frequencies is yet to be validated [75] and can be seen as a valuable future development. At the same time, the application of the method to a higher (e.g., FR-2) frequency range [189] is of interest, as the proposed VNA-based algorithm has the potential to avoid the difficulties that are encountered in phase-coherent waveform measurements at those frequencies.
The current characterization features of the WALP setup can be greatly enhanced by enabling time-domain acquisitions on the proposed VNA-based platform, where modulated envelopes and harmonics can be fully reconstructed by using phase stitching or similar methods [71]. These improvements would allow to compare the performance of the proposed VNA-based frequency-by-frequency algorithm with traditional WALP methods where full waveform measurements are available. Moreover, time-domain acquisitions can be used to develop rapid load pull characterization methods, where modulated signals are used in order to rapidly sweep an equivalent narrowband active load across a region of interest on the Smith chart [88], [231].
These capabilities would also allow to replicate, using single-transistor measurements, the broadband operating conditions encountered in load-modulated transmitters (e.g., Doherty, outphasing or LMBA) or in cross-coupled MIMO PA arrays. While the possibility of this complex emulation has already been proposed in some works [84], [85], the experimental evaluation of these prospected capabilities in the wideband case is still an active research topic.
While WALP has clearly demonstrated its capabilities in testing and reproducing realistic operating condition in order to emulate broadband mismatch [75], its effective use in PA design flows is still somewhat unclear. Indeed, one of the main goals of load pull is to help designers choose termination impedances that optimize some metrics of interest (i.e. PAE, EVM, Gain). This is straightforward and well-attested in the single-tone case, as it is feasible to brute-force synthesize many candidate loads in a given portion of the Smith

Chart and select the optimal ones, but the sheer dimensionality of the search space in typical multitone cases prevents this type of approach.
Therefore, at the current state of things, it possible to test the effect of the DUT LSOP of any broadband load, but an automatic selection procedure for optimal termination, to the best of the author's knowledge, is not yet available. A possible research development could involve the parametrization of the candidate loads by fixing a user-chosen matching network topology and expressing the profile across frequency in terms of the actual circuit components values (e.g., inductance, capacitance, transmission line's length and characteristic impedance). The numbers of such parameters is typically much lower than the number of tones in standard-emulating signals (i.e., tens vs tens-of-thousands). The low dimensionality of the resulting search space can be realistically explored using careful design-of-experiments and adaptive sampling approaches already proposed for load pull testing in the single-tone case [232], [233].

6.2.3 Design and optimization of MIMO PA models

MIMO systems are now experiencing widespread adoption in RF transmitters. Supply modulated PAs, for example, can be represented using a RF signal input, a LF supply modulation input and a single RF signal output. Similarly, dual-input Doherty and outphasing PAs can be modeled in general as dual-RF input and single-RF output systems. Large (e.g., 4x4 or more) PA arrays used in beamforming transmitters for massive MIMO 5G applications are also a relevant example. In all cases, there is significant cross-interaction between the multiple paths in the system, making its decomposition in a parallel configuration of independent SISO signal chains unfeasible.
While Volterra series, in its various implementations, has been shown to be an helpful tool in SISO RF PA modeling and DPD, its use in the MIMO case is still limited. This stems from the fact that the number and size of convolution kernels grows exponentially with the number of inputs/outputs of interest, making the approach highly unpractical. Some tailored truncations have been proposed for the task, but in many cases it is still unclear which terms should be included in the reduced expansion in order to have compact and accurate implementations [234], [235].
Therefore it would be interesting to investigate the possibility of extending Volterra-like modeling to the general MIMO case, making use of machine learning tools (e.g., kernel regression) that in the SISO case have been used to reduce the dimensionality of the problem at hand and automatically select the relevant terms in the expansion [236], [237].
The attractiveness of the Volterra approach is the use of polynomial basis functions, which are intrinsically well suited for compact representation and fast computation in simulation tools or FPGA predistorters. This is in contrast with black-box neural-network based approaches, which can also be of interest in this type of analysis [238].
On top of enabling system-level simulation and representation, these compact MIMO-PA representations are required in order to design novel multi-channel DPD architectures

[235] and as surrogate models [239] in system-level optimizations of the whole complex transmitter [240].

Bibliography

[1] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[3] 3GPP, *Technical Specification Group Radio Access Network 38.104, 38.141*, 9 2019.

[4] K. Bumman, M. Junghwan, and K. Ildu, "Efficiently amplified," *IEEE Microwave Magazine*, vol. 11, no. 5, pp. 87–100, 2010.

[5] S. Wei, D. L. Goeckel, and P. A. Kelly, "Convergence of the complex envelope of bandlimited ofdm signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4893–4904, 2010.

[6] S. C. Cripps, *RF Power Amplifiers for Wireless Communications, Second Edition (Artech House Microwave Library (Hardcover))*. USA: Artech House, Inc., 2006.

[7] J. Moon, S. Jee, J. Kim, J. Kim, and B. Kim, "Behaviors of class-F and class-F^{-1} amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 6, pp. 1937–1951, 2012.

[8] V. Camarchia, M. Pirola, R. Quaglia, S. Jee, Y. Cho, and B. Kim, "The Doherty power amplifier: Review of recent solutions and trends," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 2, pp. 559–571, 2015.

[9] D. J. Shepphard, J. Powell, and S. C. Cripps, "A broadband reconfigurable load modulated balanced amplifier (LMBA)," in *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 947–949, June 2017.

[10] T. Cappello, T. W. Barton, C. Florian, M. Litchfield, and Z. Popovic, "Multilevel supply-modulated Chireix outphasing with continuous input modulation," *IEEE*

Transactions on Microwave Theory and Techniques, vol. 65, no. 12, pp. 5231–5243, 2017.

[11] C. Florian, T. Cappello, A. Santarelli, D. Niessen, F. Filicori, and Z. Popović, "A prepulsing technique for the characterization of GaN power amplifiers with dynamic supply under controlled thermal and trapping states," *IEEE Trans. Microw. Theory Techn.*, vol. 65, pp. 5046–5062, Dec. 2017.

[12] J. Jeong, D. F. Kimball, M. Kwak, P. Draxler, C. Hsia, C. Steinbeiser, T. Landon, O. Krutko, L. E. Larson, and P. M. Asbeck, "High-efficiency WCDMA envelope tracking base-station amplifier implemented with GaAs HVHBTs," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 10, pp. 2629–2639, 2009.

[13] C. Florian, T. Cappello, R. P. Paganelli, D. Niessen, and F. Filicori, "Envelope tracking of an RF high power amplifier with an 8-level digitally controlled GaN-on-Si supply modulator," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 8, pp. 2589–2602, 2015.

[14] G. P. Gibiino, A. Santarelli, D. Schreurs, and F. Filicori, "Two-input nonlinear dynamic model inversion for the linearization of envelope-tracking RF PAs," *IEEE Microwave and Wireless Components Letters*, vol. 27, no. 1, pp. 79–81, 2016.

[15] L. R. Kahn, "Single-sideband transmission by envelope elimination and restoration," *Proceedings of the IRE*, vol. 40, no. 7, pp. 803–806, 1952.

[16] H. Wang *et al.*, "Power amplifiers performance survey 2000-present," 2020. https://gems.ece.gatech.edu/PA_survey.html, [Online].

[17] H. Amano *et al.*, "The 2018 GaN power electronics roadmap," *Journal of Physics D: Applied Physics*, vol. 51, p. 163001, Mar. 2018.

[18] F. Costanzo, R. Giofrè, A. Salvucci, G. Polli, and E. Limiti, "A 4W 37.5-42.5 GHz power amplifier MMIC in GaN on Si technology," in *Proc. Conf. Ph.D. Research in Microelectronics and Electronics (PRIME)*, pp. 137–140, July 2018.

[19] U. K. Mishra, L. Shen, T. E. Kazior, and Y. Wu, "GaN-based RF power devices and amplifiers," *Proceedings of the IEEE*, vol. 96, no. 2, pp. 287–305, 2008.

[20] T. Mimura, "Development of high electron mobility transistor," *Japanese Journal of Applied Physics*, vol. 44, pp. 8263–8268, Dec. 2005.

[21] K. Gurnett and T. Adams, "Native substrates for GaN: the plot thickens," *III-Vs Review*, vol. 19, no. 9, pp. 39–41, 2006.

[22] N. K. Subramani, "Physics-based TCAD device simulations and measurements of GaN HEMT technology for RF power devices," *PhD Thesis - Université de Limoges - École Doctorale Sciences et Ingénierie pour l 'Information*, 2018.

[23] F. Medjdoub, M. Zegaoui, N. Rolland, and P. A. Rolland, "Demonstration of low leakage current and high polarization in ultrathin AlN/GaN high electron mobility transistors grown on silicon substrate," *Applied Physics Letters*, vol. 98, no. 22, p. 223502, 2011.

[24] V. Putcha, E. Bury, J. Franco, A. Walke, S. E. Zhao, U. Peralagu, M. Zhao, A. Alian, B. Kaczer, N. Waldron, D. Linten, B. Parvais, and N. Collaert, "Exploring the DC reliability metrics for scaled GaN-on-Si devices targeted for RF/5G applications," in *2020 IEEE International Reliability Physics Symposium (IRPS)*, pp. 1–8, 2020.

[25] R. J. Trew, G. L. Bilbro, W. Kuang, Y. Liu, and H. Yin, "Microwave AlGaN/GaN HFETs," *IEEE Microwave Magazine*, vol. 6, no. 1, pp. 56–66, 2005.

[26] J. P. Ibbetson, P. T. Fini, K. D. Ness, S. P. DenBaars, J. S. Speck, and U. K. Mishra, "Polarization effects, surface states, and the source of electrons in AlGaN/GaN heterostructure field effect transistors," *Applied Physics Letters*, vol. 77, no. 2, pp. 250–252, 2000.

[27] H. Chandrasekar, K. N. Bhat, M. Rangarajan, S. Raghavan, and N. Bhat, "Thickness dependent parasitic channel formation at AlN/Si interfaces," *Scientific Reports*, vol. 7, p. 15749, Nov 2017.

[28] M. Bouslama, V. Gillet, C. Chang, J. Nallatamby, R. Sommet, M. Prigent, R. Quéré, and B. Lambert, "Dynamic performance and characterization of traps using different measurements techniques for the new AlGaN/GaN HEMT of 0.15-μm ultrashort gate length," *IEEE Trans. Microw. Theory Techn.*, pp. 1–8, 2019.

[29] C. F. Campbell, M.-Y. Kao, and S. Nayak, "High efficiency Ka-band power amplifier MMICs fabricated with a 0.15 μm GaN on SiC HEMT process," in *2012 IEEE/MTT-S International Microwave Symposium Digest*, pp. 1–3, IEEE, 2012.

[30] F. Medjdoub, M. Zegaoui, B. Grimbert, D. Ducatteau, N. Rolland, and P. Rolland, "First demonstration of high-power GaN-on-silicon transistors at 40 ghz," *IEEE electron device letters*, vol. 33, no. 8, pp. 1168–1170, 2012.

[31] T. Kazior, J. LaRoche, and W. Hoke, "More than moore: GaN HEMTs and Si CMOS get it together," in *2013 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, pp. 1–4, IEEE, 2013.

[32] A. Gasmi, M. El Kaamouchi, J. Poulain, B. Wroblewski, F. Lecourt, G. Dagher, P. Frijlink, and R. Leblanc, "10W power amplifier and 3W transmit/receive module with 3 dB NF in Ka band using a 100nm GaN/Si process," in *2017 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, pp. 1–4, IEEE, 2017.

[33] F. Iucolano and T. Boles, "GaN-on-Si HEMTs for wireless base stations," *Materials Science in Semiconductor Processing*, vol. 98, pp. 100–105, 2019.

[34] U. Peralagu, A. Alian, V. Putcha, A. Khaled, R. Rodriguez, A. Sibaja-Hernandez, S. Chang, E. Simoen, S. Zhao, B. De Jaeger, *et al.*, "CMOS-compatible GaN-based devices on 200mm-Si for RF applications: Integration and performance," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 17–2, IEEE, 2019.

[35] M. Whiteside, S. Arulkumaran, Y. Dikme, A. Sandupatla, and G. I. Ng, "Demonstration of AlGaN/GaN MISHEMT on Si with low-temperature epitaxy grown AlN dielectric gate," *Electronics*, vol. 9, no. 11, p. 1858, 2020.

[36] S. Arulkumaran, T. Egawa, H. Ishikawa, and T. Jimbo, "High-temperature effects of AlGaN/GaN high-electron-mobility transistors on sapphire and semi-insulating SiC substrates," *Applied Physics Letters*, vol. 80, no. 12, pp. 2186–2188, 2002.

[37] D. Francis, F. Faili, D. Babić, F. Ejeckam, A. Nurmikko, and H. Maris, "Formation and characterization of 4-inch GaN-on-diamond substrates," *Diamond and Related Materials*, vol. 19, no. 2, pp. 229–233, 2010. NDNC 2009.

[38] M. J. Uren, K. J. Nash, R. S. Balmer, T. Martin, E. Morvan, N. Caillas, S. L. Delage, D. Ducatteau, B. Grimbert, and J. C. De Jaeger, "Punch-through in short-channel AlGaN/GaN HFETs," *IEEE Transactions on Electron Devices*, vol. 53, pp. 395–398, Feb. 2006.

[39] L. Ravikiran, K. Radhakrishnan, S. Munawar Basha, N. Dharmarasu, M. Agrawal, C. M. Manoj kumar, S. Arulkumaran, and G. I. Ng, "Study on GaN buffer leakage current in AlGaN/GaN high electron mobility transistor structures grown by ammonia-molecular beam epitaxy on 100-mm Si(111)," *Journal of Applied Physics*, vol. 117, no. 24, p. 245305, 2015.

[40] J. Bergtsen, "Buffer related dispersive effects in microwave GaN HEMTs," *PhD Thesis - Chalmers University of Technology - Department of Microtechnology and Nanoscience*, 2018.

[41] S. Gustafsson, J. Chen, J. Bergsten, U. Forsberg, M. Thorsell, E. Janzén, and N. Rorsman, "Dispersive effects in microwave AlGaN/AlN/GaN HEMTs with carbon-doped buffer," *IEEE Transactions on Electron Devices*, vol. 62, no. 7, pp. 2162–2169, 2015.

[42] V. Desmaris, M. Rudzinski, N. Rorsman, P. R. Hageman, P. K. Larsen, H. Zirath, T. C. Rodle, and H. F. F. Jos, "Comparison of the DC and microwave performance of AlGaN/GaN HEMTs grown on SiC by MOCVD with Fe-doped or unintentionally doped GaN buffer layers," *IEEE Transactions on Electron Devices*, vol. 53, no. 9, pp. 2413–2417, 2006.

[43] K. Sharma, E. Dupouy, M. Bouslama, R. Sommet, and J. C. Nallatamby, "Impact of the location of iron buffer doping on trap signatures in GaN HEMTs," in *2020 International Workshop on Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMiC)*, pp. 1–3, 2020.

[44] D. Chen, A. Malmros, M. Thorsell, H. Hjelmgren, O. Kordina, J. Chen, and N. Rorsman, "Microwave performance of 'buffer-free' GaN-on-SiC high electron mobility transistors," *IEEE Electron Device Letters*, vol. 41, no. 6, pp. 828–831, 2020.

[45] R. Kabouche, J. Derluyn, R. Pusche, S. Degroote, M. Germain, R. Pecheux, E. Okada, M. Zegaoui, and F. Medjdoub, "Comparison of C-doped AlN/GaN HEMTs and AlN/GaN/AlGaN double heterostructure for mmW applications," in *2018 13th European Microwave Integrated Circuits Conference (EuMIC)*, pp. 5–8, Sept. 2018.

[46] R. Pecheux, R. Kabouche, E. Dogmus, A. Linge, E. Okada, M. Zegaoui, and F. Medjdoub, "Importance of buffer configuration in GaN HEMTs for high microwave performance and robustness," in *2017 47th European Solid-State Device Research Conference (ESSDERC)*, pp. 228–231, Sept. 2017.

[47] D. Bisi, M. Meneghini, C. de Santi, A. Chini, M. Dammann, P. Brückner, M. Mikulla, G. Meneghesso, and E. Zanoni, "Deep-level characterization in GaN HEMTs-Part I: Advantages and limitations of drain current transient measurements," *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3166–3175, 2013.

[48] S. A. Albahrani, A. E. Parker, and M. Heimlich, "Identifying a double-energy-level trap center in a gan hemt by performing three-stage pulse measurements," *IEEE Transactions on Electron Devices*, vol. 63, no. 9, pp. 3693–3699, 2016.

[49] J. G. Rathmell and A. E. Parker, "Circuit implementation of a theoretical model of trap centres in GaAs and GaN devices," in *Microelectronics: Design, Technology, and Packaging III* (A. J. Hariz and V. K. Varadan, eds.), vol. 6798, pp. 187 – 197, International Society for Optics and Photonics, SPIE, 2007.

[50] J. L. Gomes, L. C. Nunes, and J. C. Pedro, "Explaining the different time constants extracted from low frequency y22 and i_{DS}-DLTS on GaN HEMTs," in *2020 IEEE/MTT-S International Microwave Symposium (IMS)*, pp. 432–435, 2020.

[51] Chih-Tang Sah, "The equivalent circuit model in solid-state electronics—part i: The single energy level defect centers," *Proceedings of the IEEE*, vol. 55, no. 5, pp. 654–671, 1967.

[52] D. Bisi, "Characterization of charge trapping phenomena in GaN-based HEMTs," *PhD Thesis - Universitá di Padova*, 2015.

[53] S. A. Albahrani, A. Parker, and M. Heimlich, "Circuit model for double-energy-level trap centers in GaN HEMTs," *IEEE Trans. Electron Devices*, vol. 64, no. 3, pp. 998–1006, 2017.

[54] A. M. Angelotti, G. P. Gibiino, A. Santarelli, and C. Florian, "Experimental characterization of charge trapping dynamics in 100-nm AlN/GaN/AlGaN-on-Si HEMTs by wideband transient measurements," *IEEE Transactions on Electron Devices*, vol. 67, no. 8, pp. 3069–3074, 2020.

[55] J. Joh and J. A. del Alamo, "RF power degradation of GaN high electron mobility transistors," in *2010 International Electron Devices Meeting*, pp. 20.2.1–20.2.4, Dec. 2010.

[56] R. Vetury, N. Q. Zhang, S. Keller, and U. K. Mishra, "The impact of surface states on the DC and RF characteristics of AlGaN/GaN HFETs," *IEEE Trans. Electron Devices*, vol. 48, pp. 560–566, Mar. 2001.

[57] H. Kim, R. M. Thompson, V. Tilak, T. R. Prunty, J. R. Shealy, and L. F. Eastman, "Effects of SiN passivation and high-electric field on AlGaN-GaN HFET degradation," *IEEE Electron device letters*, vol. 24, no. 7, pp. 421–423, 2003.

[58] H. Huang, Y. C. Liang, G. S. Samudra, T.-F. Chang, and C.-F. Huang, "Effects of gate field plates on the surface state related current collapse in AlGaN/GaN HEMTs," *IEEE transactions on power electronics*, vol. 29, no. 5, pp. 2164–2173, 2013.

[59] A. Benvegnu, S. Laurent, M. Meneghini, G. Meneghesso, J. L. Muraro, D. Barataud, E. Zanoni, and R. Quere, "Trap characterization of AlGaN/GaN HEMTs through drain current measurements under pulsed-RF large-signal excitation," *2015 IEEE MTT-S Int. Microw. Symp. IMS 2015*, pp. 1–4, 2015.

[60] O. Jardel, F. De Groote, C. Charbonniaud, T. Reveyrand, J. P. Teyssier, R. Quere, and D. Floriot, "A drain-lag model for AlGaN/GaN power HEMTs," in *2007 IEEE/MTT-S International Microwave Symposium*, pp. 601–604, 2007.

[61] D. Williams, "Traveling waves and power waves: Building a solid foundation for microwave circuit theory," *IEEE Microwave Magazine*, vol. 14, no. 7, pp. 38–45, 2013.

[62] R. Marks and D. Williams, "A general waveguide circuit theory," 1992-10-01 1992.

[63] R. Pintelon and J. Schoukens, *System identification: a frequency domain approach*. John Wiley & Sons, 2001.

[64] N. B. Carvalho, K. A. Remley, D. Schreurs, and K. G. Gard, "Multisine signals for wireless system test and design [application notes]," *IEEE Microw. Mag.*, vol. 9, pp. 122–138, June 2008.

[65] C. Nader, W. Van Moer, N. Bjorsell, and P. Handel, "Wideband radio frequency measurements: From instrumentation to sampling theory," *IEEE Microwave Magazine*, vol. 14, no. 2, pp. 85–98, 2013.

[66] S. Gustafsson, M. Thorsell, J. Stenarson, and C. Fager, "An oscilloscope correction method for vector-corrected RF measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 9, pp. 2541–2547, 2015.

[67] D. Williams, P. Hale, and K. A. Remley, "The sampling oscilloscope as a microwave instrument," *IEEE Microwave Magazine*, vol. 8, no. 4, pp. 59–68, 2007.

[68] *Modern RF and Microwave Measurement Techniques*. The Cambridge RF and Microwave Engineering Series, Cambridge University Press, 2013.

[69] G. P. Gibiino, A. M. Angelotti, A. Santarelli, and C. Florian, "Microwave characterization of trapping effects in 100-nm GaN-on-Si HEMT technology," *IEEE Microwave and Wireless Components Letters*, vol. 29, no. 9, pp. 604–606, 2019.

[70] K. Anderson, J. Verspecht, and T. S. Nielsen, "Method and system for preventing interference caused by mirror frequencies," Patent - US9793857B1.

[71] T. S. Nielsen, J. Verspecht, and J. P. Dunsmore, "Method and apparatus for spectral stitching using reference channel and pilot tones," U.S. Patent - US20170047915A1.

[72] *VNA Measurement Systems*, ch. 2, pp. 66–123. John Wiley & Sons, Ltd, 2012.

[73] J. Verspecht, J. P. Teyssier, and T. S. Nielsen, "Method and system for preventing interference caused by mirror frequencies," Patent - DE102018208465A1.

[74] A. M. Angelotti, G. P. Gibiino, T. Nielsen, F. F. Tafuri, and A. Santarelli, "Three port non-linear characterization of power amplifiers under modulated excitations using a vector network analyzer platform," in *2018 IEEE/MTT-S International Microwave Symposium - IMS*, pp. 1021–1024, 2018.

[75] M. Marchetti, M. J. Pelk, K. Buisman, W. C. E. Neo, M. Spirito, and L. C. N. de Vreede, "Active harmonic load–pull with realistic wideband communications signals," *IEEE Trans. Microw. Theory Techn.*, vol. 56, pp. 2979–2988, Dec 2008.

[76] and P. Roblin, , , and B. Poling, "New thermometry and trap relaxation characterization techniques for AlGaN/GaN HEMTs using pulsed-RF excitations," in *2012 IEEE/MTT-S International Microwave Symposium Digest*, pp. 1–3, June 2012.

[77] *Advanced Measurement Techniques*, ch. 9, pp. 552–599. John Wiley & Sons, Ltd, 2012.

[78] P. Roblin, *Nonlinear RF Circuits and Nonlinear Vector Network Analyzers: Interactive Measurement and Design Techniques*. The Cambridge RF and Microwave Engineering Series, Cambridge University Press, 2011.

[79] Y. Zhang, Z. He, H. Li, and M. Nie, "Dense spectral grid NVNA phase measurements using vector signal generators," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 12, pp. 2983–2992, 2014.

[80] D. Niessen *et al.*, "Charge-controlled GaN FET modeling by displacement current integration from frequency-domain NVNA measurements," *IEEE Trans. Microw. Theory Techn.*, vol. 64, pp. 4382–4393, Dec. 2016.

[81] F. Verbeyst and M. V. Bossche, "The Volterra input-output map of a high frequency amplifier as a practical alternative to load-pull measurements," in *Conference Proceedings. 10th Anniversary. IMTC/94. Advanced Technologies in I M. 1994 IEEE Instrumentation and Measurement Technolgy Conference (Cat. No.94CH3424-9)*, pp. 283–286 vol.1, 1994.

[82] A. Ferrero and M. Pirola, "Harmonic load-pull techniques: An overview of modern systems," *IEEE Microwave Magazine*, vol. 14, no. 4, pp. 116–123, 2013.

[83] M. S. Hashmi and F. M. Ghannouchi, "Introduction to load-pull systems and their applications," *IEEE Instrumentation Measurement Magazine*, vol. 16, no. 1, pp. 30–36, 2013.

[84] W. Hallberg, D. Nopchinda, C. Fager, and K. Buisman, "Emulation of doherty amplifiers using single-amplifier load–pull measurements," *IEEE Microw. Wirel. Compon. Lett.*, vol. 30, no. 1, pp. 47–49, 2019.

[85] D. Nopchinda and K. Buisman, "Measurement technique to emulate signal coupling between power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 4, pp. 2034–2046, 2018.

[86] G. P. Gibiino, K. Łukasik, P. Barmuta, A. Santarelli, D. M.-P. Schreurs, and F. Filicori, "A two-port nonlinear dynamic behavioral model of RF PAs subject to wideband load modulation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 2, pp. 831–844, 2017.

[87] D. Niessen, G. P. Gibiino, R. Cignani, A. Santarelli, D. M. M. . Schreurs, and F. Filicori, "Charge-controlled GaN FET modeling by displacement current integration from frequency-domain NVNA measurements," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 12, pp. 4382–4393, 2016.

[88] F. Verbeyst and M. V. Bossche, "Real-time and optimal pa characterization speeds up pa design," in *34th European Microwave Conference, 2004.*, vol. 1, pp. 431–434, 2004.

[89] K. Lukasik, P. Barmuta, T. Nielsen, K. Madziar, D. Schreurs, and A. Lewandowski, "Hybrid load-pull system using a two-source nonlinear vector network analyzer," in *2015 Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC)*, pp. 1–3, 2015.

[90] S. Khandelwal *et al.*, "ASM GaN: Industry standard model for GaN RF and power devices—part 1: DC, CV, and RF model," *IEEE Trans. Electron Devices*, vol. 66, pp. 80–86, Jan. 2019.

[91] N. K. Subramani, J. Couvidat, A. Al Hajjar, J.-C. Nallatamby, R. Sommet, and R. Quéré, "Identification of GaN buffer traps in microwave power AlGaN/GaN HEMTs through low frequency S-parameters measurements and TCAD-based physical device simulations," *IEEE Journal of the Electron Devices Society*, vol. 5, no. 3, pp. 175–181, 2017.

[92] J. Xu, R. Jones, S. A. Harris, T. Nielsen, and D. E. Root, "Dynamic FET model - DynaFET - for GaN transistors from NVNA active source injection measurements," in *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*, pp. 1–3, 2014.

[93] F. Filicori, G. Vannini, and V. A. Monaco, "A nonlinear integral model of electron devices for HB circuit analysis," *IEEE Transactions on Microwave Theory and Techniques*, vol. 40, no. 7, pp. 1456–1465, 1992.

[94] J. G. Leckey, "A scalable X-parameter model for GaAs and GaN FETs," in *2011 6th European Microwave Integrated Circuit Conference*, pp. 13–16, 2011.

[95] I. Angelov, L. Bengtsson, and M. Garcia, "Extensions of the Chalmers nonlinear HEMT and MESFET model," *IEEE Transactions on Microwave Theory and Techniques*, vol. 44, no. 10, pp. 1664–1674, 1996.

[96] D. Mirri, F. Filicori, G. Iuculano, and G. Pasini, "A nonlinear dynamic model for performance analysis of large-signal amplifiers in communication systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 2, pp. 341–350, 2004.

[97] J. Verspecht and D. E. Root, "Polyharmonic distortion modeling," *IEEE Microw. Mag.*, vol. 7, no. 3, pp. 44–57, 2006.

[98] G. P. Gibiino, G. Avolio, D. M.-P. Schreurs, A. Santarelli, and F. Filicori, "A three-port nonlinear dynamic behavioral model for supply-modulated RF PAs," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 1, pp. 133–147, 2015.

[99] S. D. and al., *RF Power Amplifier Behavioral Modeling*. The Cambridge RF and Microwave Engineering Series, Cambridge University Press, 2008.

[100] N. B. De Carvalho and J. C. Pedro, "Large- and small-signal IMD behavior of microwave power amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 47, no. 12, pp. 2364–2374, 1999.

[101] J. Verspecht, J. Horn, L. Betts, D. Gunyan, R. Pollard, C. Gillease, and D. E. Root, "Extension of X-parameters to include long-term dynamic memory effects," in *2009 IEEE MTT-S International Microwave Symposium Digest*, pp. 741–744, 2009.

[102] J. Couvidat, N. K. Subramani, V. Gillet, S. Laurent, C. Charbonniaud, J. C. Nallatamby, M. Prigent, N. Deltimple, and R. Quere, "Investigation of fast and slow charge trapping mechanisms of GaN/AlGaN HEMTs through pulsed I-V measurements and the associated new trap model," in *2018 IEEE/MTT-S International Microwave Symposium - IMS*, pp. 720–723, June 2018.

[103] L. C. Nunes *et al.*, "A simple method to extract trapping time constants of GaN HEMTs," in *IEEE/MTT-S Int. Microw. Symp. Dig.*, pp. 716–719, June 2018.

[104] J. C. Pedro and S. A. Maas, "A comparative overview of microwave and wireless power-amplifier behavioral modeling approaches," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 4, pp. 1150–1163, 2005.

[105] S. Boyd and L. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Transactions on Circuits and Systems*, vol. 32, no. 11, pp. 1150–1161, 1985.

[106] C. Crespo-Cadenas, J. Reina-Tosina, and M. J. Madero-Ayora, "Volterra behavioral model for wideband RF amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 55, no. 3, pp. 449–457, 2007.

[107] A. Santarelli *et al.*, "Multi-bias nonlinear characterization of GaN FET trapping effects through a multiple pulse time domain network analyzer," in *Proc. European Microw. Int. Circ. Conf. (EuMIC)*, pp. 81–84, Sept. 2015.

[108] G. Avolio *et al.*, "Dynamic-bias S-parameters: A new measurement technique for microwave transistors," *IEEE Trans. Microw. Theory Techn.*, vol. 64, pp. 3946–3955, Nov. 2016.

[109] S. Boyd, Y. Tang, and L. Chua, "Measuring Volterra kernels," *IEEE Transactions on Circuits and Systems*, vol. 30, no. 8, pp. 571–577, 1983.

[110] D. Mirri, G. Iuculano, F. Filicori, G. Vannini, G. Pasini, and G. Pellegrini, "A modified Volterra series approach for the characterization of non-linear dynamic systems," in *Quality Measurement: The Indispensable Bridge between Theory and Reality (No Measurements? No Science! Joint Conference-1996: IEEE Instrumentation and Measurement Technology Conference and IMEKO Tec*, vol. 1, pp. 710–715, IEEE Instrumentation and Measurement Technology Conference, 1996.

[111] A. Cooman, P. Bronders, D. Peumans, G. Vandersteen, and Y. Rolain, "Distortion contribution analysis with the best linear approximation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, pp. 4133–4146, Dec 2018.

[112] Y. Rolain, W. Van Moer, R. Pintelon, and J. Schoukens, "Experimental characterization of the nonlinear behavior of rf amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 8, pp. 3209–3218, 2006.

[113] J. Verspecht, A. Stav, J. Teyssier, and S. Kusano, "Characterizing amplifier modulation distortion using a vector network analyzer," in *ARFTG Microw. Meas. Conf. Dig.*, pp. 1–4, June 2019.

[114] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*. John Wiley and Sons, 2012.

[115] C. P. Silva, C. J. Clark, A. A. Moulthrop, and M. S. Muha, "Optimal-filter approach for nonlinear power amplifier modeling and equalization," in *2000 IEEE MTT-S International Microwave Symposium Digest (Cat. No.00CH37017)*, vol. 1, pp. 437–440 vol.1, 2000.

[116] Y. Rolain, M. Zyari, E. Van Nechel, and G. Vandersteen, "A measurement-based error-vector-magnitude model to assess non linearity at the system level," in *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 1429–1432, June 2017.

[117] K. Freiberger, H. Enzinger, and C. Vogel, "A noise power ratio measurement method for accurate estimation of the error vector magnitude," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 5, pp. 1632–1645, 2017.

[118] A. Cooman, P. Bronders, D. Peumans, G. Vandersteen, and Y. Rolain, "Distortion contribution analysis with the best linear approximation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, pp. 4133–4146, Dec. 2018.

[119] M. D. Mckinley, K. A. Remley, M. Myslinski, J. S. Kenney, D. Schreurs, and B. Nauwelaers, "EVM calculation for broadband modulated signals," in *ARFTG Microw. Meas. Conf. Dig., 2004*, 2004.

[120] D. Floriot, V. Brunel, M. Camiade, C. Chang, B. Lambert, Z. Ouarch-Provost, H. Blanck, J. Grünenpütt, M. Hosch, H. Jung, *et al.*, "GH25-10: New qualified power GaN HEMT process from technology to product overview," in *European Microwave Integrated Circuit Conference*, pp. 225–228, 2014.

[121] V. Di Giacomo-Brunel, E. Byk, C. Chang, J. Grünenpütt, B. Lambert, G. Mouginot, D. Sommer, H. Jung, M. Camiade, P. Fellon, *et al.*, "Industrial 0.15-μm algan/gan on SiC technology for applications up to Ka band," in *2018 13th European Microwave Integrated Circuits Conference (EuMIC)*, pp. 1–4, IEEE, 2018.

[122] M. Bouslama, A. Al Hajjar, R. Sommet, F. Medjdoub, and J.-C. Nallatamby, "Characterization and electrical modeling including trapping effects of AlN/GaN hemt 4×50 μm on silicon substrate," in *European Microwave Conference (EuMC)*, pp. 1301–1304, 2018.

[123] M. Bouslama, V. Gillet, C. Chang, J.-C. Nallatamby, R. Sommet, M. Prigent, R. Quéré, and B. Lambert, "Dynamic performance and characterization of traps using different measurements techniques for the new AlGaN/GaN HEMT of 0.15-μm ultrashort gate length," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 7, pp. 2475–2482, 2019.

[124] C. Florian, T. Cappello, A. Santarelli, D. Niessen, F. Filicori, and Z. Popović, "A prepulsing technique for the characterization of GaN power amplifiers with dynamic supply under controlled thermal and trapping states," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 12, pp. 5046–5062, 2017.

[125] K. Kellogg, S. Khandelwal, L. Dunleavy, and J. Wang, "Characterization of thermal and trapping time constants in a GaN HEMT," in *ARFTG Microwave Measurement Symposium (ARFTG)*, pp. 1–4, IEEE, 2020.

[126] J. Couvidat *et al.*, "Investigation of fast and slow charge trapping mechanisms of GaN/AlGaN HEMTs through pulsed I-V measurements and the associated new trap model," in *IEEE/MTT-S Int. Microw. Symp. Dig.*, pp. 720–723, June 2018.

[127] M. Uren, K. Nash, R. S. Balmer, T. Martin, E. Morvan, N. Caillas, S. L. Delage, D. Ducatteau, B. Grimbert, and J. C. De Jaeger, "Punch-through in short-channel algan/gan hfets," *IEEE Transactions on Electron Devices*, vol. 53, no. 2, pp. 395–398, 2006.

[128] M. Allaei, M. Shalchian, and F. Jazaeri, "Modeling of short-channel effects in GaN HEMTs," *IEEE Transactions on Electron Devices*, vol. 67, no. 8, pp. 3088–3094, 2020.

[129] J. Nallatamby, R. Sommet, O. Jardel, S. Laurent, M. Prigent, and R. Quere, "A microwave modeling oxymoron?: Low-frequency measurements for microwave device modeling," *IEEE Microwave Magazine*, vol. 15, no. 4, pp. 92–107, 2014.

[130] A. M. Angelotti, G. P. Gibiino, C. Florian, and A. Santarelli, "Trapping dynamics in GaN HEMTs for millimeter-wave applications: Measurement-based characterization and technology comparison," *Electronics*, vol. 10, no. 2, 2021.

[131] F. Medjdoub, D. Ducatteau, M. Zegaoui, B. Grimbert, N. Rolland, and P.-A. Rolland, "Trapping effects dependence on electron confinement in ultrashort GaN-on-si high-electron-mobility transistors," *Applied Physics Express*, vol. 5, p. 034103, Mar. 2012.

[132] G. Meneghesso, M. Meneghini, and E. Zanoni, "Breakdown mechanisms in AlGaN/GaN HEMTs: An overview," *Japanese Journal of Applied Physics*, vol. 53, p. 100211, Sept. 2014.

[133] J. Scott, J. G. Rathmell, A. Parker, and M. Sayed, "Pulsed device measurements and applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 44, no. 12, pp. 2718–2723, 1996.

[134] C. F. Gonçalves, L. C. Nunes, P. M. Cabral, and J. C. Pedro, "Pulsed I/V and S-parameters measurement system for isodynamic characterization of power GaN HEMT transistors," *Int. J. RF Microw. Computer-Aided Eng.*, vol. 28, no. 8, p. e21515, 2018.

[135] S. A. Albahrani, A. Parker, and M. Heimlich, "Circuit model for double-energy-level trap centers in GaN HEMTs," *IEEE Trans. Electron Devices*, vol. 64, pp. 998–1006, Mar. 2017.

[136] S. C. Binari, K. Ikossi, J. A. Roussos, W. Kruppa, H. B. Dietrich, D. D. Koleske, A. E. Wickenden, and R. L. Henry, "Trapping effects and microwave power performance in AlGaN/GaN HEMTs," *IEEE Trans. Electron Devices*, vol. 48, pp. 465–471, Mar. 2001.

[137] O. Axelsson *et al.*, "Application relevant evaluation of trapping effects in AlGaN/GaN HEMTs with Fe-doped buffer," *IEEE Trans. Electron Devices*, vol. 63, pp. 326–332, Jan. 2016.

[138] H. Hirshy, M. Singh, M. A. Casbon, R. M. Perks, M. J. Uren, T. Martin, M. Kuball, and P. J. Tasker, "Evaluation of pulsed I-V analysis as validation tool of nonlinear RF models of GaN-based HFETs," *IEEE Trans. Electron Devices*, vol. 65, no. 12, pp. 5307–5313, 2018.

[139] A. Santarelli *et al.*, "A double-pulse technique for the dynamic I/V characterization of GaN FETs," *IEEE Microw. Wirel. Comp. Lett.*, vol. 24, pp. 132–134, Feb. 2014.

[140] G. P. Gibiino *et al.*, "Isotrap pulsed *IV* characterization of GaN HEMTs for PA design," *IEEE Microw. Wirel. Compon. Lett.*, vol. 28, pp. 672–674, Aug. 2018.

[141] J. L. Gomes, L. C. Nunes, C. F. Gonçalves, and J. C. Pedro, "An accurate characterization of capture time constants in GaN HEMTs," *IEEE Trans. Microw. Theory Techn.*, vol. 67, pp. 2465–2474, July 2019.

[142] M. Bouya, N. Malbert, N. Labat, D. Carisetti, P. Perdu, J. Clément, B. Lambert, and M. Bonnet, "Analysis of traps effect on AlGaN/GaN HEMT by luminescence techniques," *Microelectronics Reliability*, vol. 48, no. 8, pp. 1366–1369, 2008. 19th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF 2008).

[143] D. V. Lang, "Deep-level transient spectroscopy: A new method to characterize traps in semiconductors," *J. Appl. Phys.*, vol. 45, p. 3023, 1974.

[144] J. Joh and J. A. Del Alamo, "A current-transient methodology for trap analysis for GaN high electron mobility transistors," *IEEE Trans. Electron Devices*, vol. 58, no. 1, pp. 132–140, 2011.

[145] A. Benvegnù, S. Laurent, O. Jardel, J. Muraro, M. Meneghini, D. Barataud, G. Meneghesso, E. Zanoni, and R. Quéré, "Characterization of defects in AlGaN/GaN HEMTs based on nonlinear microwave current transient spectroscopy," *IEEE Trans. Electron Devices*, vol. 64, pp. 2135–2141, May 2017.

[146] O. Mitrofanov and M. Manfra, "Poole-Frenkel electron emission from the traps in AlGaN/GaN transistors," *Journal of Applied Physics*, vol. 95, no. 11, pp. 6414–6419, 2004.

[147] J. Joh and J. A. del Alamo, "A current-transient methodology for trap analysis for GaN high electron mobility transistors," *IEEE Trans. Electron Devices*, vol. 58, pp. 132–140, Jan. 2011.

[148] C. Potier, J.-C. Jacquet, C. Dua, A. Martin, M. Campovecchio, M. Oualli, O. Jardel, S. Piotrowicz, S. Laurent, R. Aubry, and et al., "Highlighting trapping phenomena in microwave GaN HEMTs by low-frequency S-parameters," *International Journal of Microwave and Wireless Technologies*, vol. 7, no. 3-4, p. 287–296, 2015.

[149] A. Chini, F. Soci, M. Meneghini, G. Meneghesso, and E. Zanoni, "Deep levels characterization in GaN HEMTs—Part II: Experimental and numerical evaluation of self-heating effects on the extraction of traps activation energy," *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3176–3182, 2013.

[150] N. K. Subramani *et al.*, "Identification of GaN buffer traps in microwave power AlGaN/GaN HEMTs through low frequency S-parameters measurements and TCAD-based physical device simulations," *IEEE J. Electron Devices Soc.*, vol. 5, pp. 175–181, May 2017.

[151] L. C. Nunes, J. L. Gomes, P. M. Cabral, and J. C. Pedro, "A simple method to extract trapping time constants of GaN HEMTs," in *IEEE/MTT-S Int. Microw. Symp. Dig.*, pp. 716–719, June 2018.

[152] G. P. Gibiino, A. Santarelli, and F. Filicori, "A GaN HEMT global large-signal model including charge trapping for multibias operation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 11, pp. 4684–4697, 2018.

[153] J. Bergsten, M. Thorsell, D. Adolph, J. Chen, O. Kordina, E. . Sveinbjörnsson, and N. Rorsman, "Electron trapping in extended defects in microwave AlGaN/GaN HEMTs with carbon-doped buffers," *IEEE Transactions on Electron Devices*, vol. 65, no. 6, pp. 2446–2453, 2018.

[154] O. Mitrofanov and M. Manfra, "Dynamics of trapped charge in GaN/AlGaN/GaN high electron mobility transistors grown by plasma-assisted molecular beam epitaxy," *Applied Physics Letters*, vol. 84, no. 3, pp. 422–424, 2004.

[155] A. Benvegnù *et al.*, "Trap characterization of AlGaN/GaN HEMTs through drain current measurements under pulsed-RF large-signal excitation," in *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 1–4, May 2015.

[156] A. Santarelli *et al.*, "Multi-bias nonlinear characterization of GaN FET trapping effects through a multiple pulse time domain network analyzer," in *Proc. Eur. Microw. Int. Circ. Conf. (EuMIC)*, pp. 81–84, Sept. 2015.

[157] F. Medjdoub, M. Zegaoui, B. Grimbert, D. Ducatteau, N. Rolland, and P. A. Rolland, "First demonstration of high-power GaN-on-silicon transistors at 40 GHz," *IEEE Electron Device Letters*, vol. 33, pp. 1168–1170, Aug. 2012.

[158] F. Medjdoub, "Ultrathin barrier GaN-on-Silicon devices for millimeter wave applications," *Microelectronics Reliability*, vol. 54, no. 1, pp. 1–12, 2014.

[159] S. A. Albahrani, A. Parker, M. Heimlich, and B. Schwitter, "Iso-trapping measurement technique for characterization of self-heating in a GaN HEMT," *IEEE Trans. Electron Devices*, vol. 64, pp. 102–108, Jan. 2017.

[160] A. Chini, F. Soci, M. Meneghini, G. Meneghesso, and E. Zanoni, "Deep levels characterization in GaN HEMTs—part II: Experimental and numerical evaluation of self-heating effects on the extraction of traps activation energy," *IEEE Trans. Electron Devices*, vol. 60, no. 10, pp. 3176–3182, 2013.

[161] M. Bouslama, A. Al Hajjar, R. Sommet, F. Medjdoub, and J. Nallatamby, "Characterization and electrical modeling including trapping effects of AlN/GaN HEMT 4×50um on silicon substrate," in *Eur. Microw. Int. Circ. Conf. (EuMIC)*, pp. 333–336, Sept. 2018.

[162] A. Hierro, A. Arehart, B. Heying, M. Hansen, J. Speck, U. Mishra, S. DenBaars, and S. Ringel, "Capture kinetics of electron traps in MBE-grown n-GaN," *physica status solidi (b)*, vol. 228, no. 1, pp. 309–313, 2001.

[163] A. Santarelli, R. Cignani, G. P. Gibiino, D. Niessen, P. A. Traverso, C. Florian, D. M.-P. Schreurs, and F. Filicori, "A double-pulse technique for the dynamic I/V characterization of GaN FETs," *IEEE Microwave and Wireless Components Letters*, vol. 24, no. 2, pp. 132–134, 2013.

[164] J. L. Gomes, L. C. Nunes, C. F. Goncalves, and J. C. Pedro, "An accurate characterization of capture time constants in GaN HEMTs," *IEEE Trans. Microw. Theory Tech.*, vol. 67, no. 7, pp. 2465–2474, 2019.

[165] S. A. Albahrani, A. Parker, G. Town, M. Heimlich, B. Schwitter, and S. Mahon, "Characterization of trapping in a GaN HEMT by performing isothermal three-stage pulse measurements," in *European Microwave Integrated Circuits Conference (EuMIC)*, pp. 161–164, IEEE, 2016.

[166] R. Vetury, N. Q. Zhang, S. Keller, and U. K. Mishra, "The impact of surface states on the DC and RF characteristics of AlGaN/GaN HFETs," *IEEE Transactions on Electron Devices*, vol. 48, no. 3, pp. 560–566, 2001.

[167] G. P. Gibiino, C. Florian, A. Santarelli, T. Cappello, and Z. Popović, "Isotrap pulsed *iv* characterization of GaN HEMTs for PA design," *IEEE Microwave and Wireless Components Letters*, vol. 28, no. 8, pp. 672–674, 2018.

[168] C. Florian, G. P. Gibiino, and A. Santarelli, "Characterization and modeling of RF GaN switches accounting for trap-induced degradation under operating regimes," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 12, pp. 5491–5500, 2018.

[169] T. Cappello, C. Florian, A. Santarelli, and Z. Popovic, "Linearization of a 500-w L-band GaN Doherty power amplifier by dual-pulse trap characterization," in *2019 IEEE MTT-S International Microwave Symposium (IMS)*, pp. 905–908, 2019.

[170] C. Florian, T. Cappello, A. Santarelli, D. Niessen, F. Filicori, and Z. Popović, "A prepulsing technique for the characterization of GaN power amplifiers with dynamic supply under controlled thermal and trapping states," *IEEE Trans. Microw. Theory Techn.*, vol. 65, pp. 5046–5062, Dec. 2017.

[171] H. Hirshy, M. Singh, M. A. Casbon, R. M. Perks, M. J. Uren, T. Martin, M. Kuball, and P. J. Tasker, "Evaluation of pulsed I–V analysis as validation tool of nonlinear RF models of GaN-Based HFETs," *IEEE Trans. Electron Devices*, vol. 65, pp. 5307–5313, Dec. 2018.

[172] A. M. Angelotti, G. P. Gibiino, C. Florian, and A. Santarelli, "Narrow-pulse-width double-pulsed S-parameters measurements of 100-nm GaN-on-Si HEMTs," in *2019 14th European Microwave Integrated Circuits Conference (EuMIC)*, pp. 17–20, 2019.

[173] A. M. Angelotti, G. P. Gibiino, C. Florian, and A. Santarelli, "Broadband error vector magnitude characterization of a GaN power amplifier using a vector network analyzer," in *2020 IEEE/MTT-S International Microwave Symposium (IMS)*, pp. 747–750, 2020.

[174] G. P. Gibiino, A. Santarelli, and F. Filicori, "Charge-conservative GaN HEMT nonlinear modeling from non-isodynamic multi-bias S-parameter measurements," *International Journal of Microwave and Wireless Technologies*, vol. 11, no. 5-6, p. 431–440, 2019.

[175] C. F. Gonçalves, L. C. Nunes, P. M. Cabral, and J. C. Pedro, "Conservative current and charge data extracted from pulsed S-parameter measurements for GaN HEMT PA design," in *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 1065–1068, June 2017.

[176] Application Note 1408-12, Keysight Technologies, *Pulsed-RF S-Parameter Measurements with the PNA Microwave Network Analyzers Using Wideband and Narrowband Detection*, May 2006.

[177] A. E. Rafei *et al.*, "DC (10 Hz) to RF (40 GHz) output conduction extraction by S-parameters measurements for in-depth characterization of AlInN/GaN HEMTs, focusing on low frequency dispersion effects," in *Proc. European Microw. Int. Circ. Conf. (EuMIC)*, pp. 5–8, Oct. 2011.

[178] G. P. Gibiino *et al.*, "Active baseband drain-supply terminal load-pull of an X-band GaN MMIC PA," in *Proc. European Microw. Conf. (EuMC)*, pp. 1204–1207, Sept. 2015.

[179] V. Gillet, M. Bouslama, J. Teyssier, M. Prigent, J. Nallatamby, and R. Quéré, "An unequally spaced multi-tone load–pull characterization technique for simultaneous linearity and efficiency assessment of RF power devices," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, pp. 2505–2513, July 2019.

[180] R. Figueiredo, A. Piacibello, V. Camarchia, and N. B. Carvalho, "Swept notch NPR for linearity assessment of systems presenting long-term memory effects," in *2020 95th ARFTG Microwave Measurement Conference (ARFTG)*, pp. 1–4, 2020.

[181] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Transactions on Signal Processing*, vol. 54, pp. 3852–3860, Oct. 2006.

[182] H. Arthaber, M. L. Mayer, and G. Magerl, "A broadband active harmonic load-pull setup with a modulated generator as active load," in *Proc. Eur. Microw. Conf.*, vol. 2, pp. 685–688, Oct 2004.

[183] S. Gustafsson, M. Thorsell, and C. Fager, "A novel active load-pull system with multi-band capabilities," in *ARFTG Microw. Meas. Conf.*, pp. 1–4, June 2013.

[184] S. Alsahali *et al.*, "A novel modulated rapid load pull system with digital predistortion capabilities," in *ARFTG Microw. Meas. Conf.*, pp. 1–4, June 2019.

[185] A. M. Angelotti, G. P. Gibiino, T. S. Nielsen, D. Schreurs, and A. Santarelli, "Enhanced wideband active load-pull with a vector network analyzer using modulated excitations and device output match compensation," in *2020 IEEE/MTT-S International Microwave Symposium (IMS)*, pp. 763–766, 2020.

[186] A. M. Angelotti, G. P. Gibiino, T. S. Nielsen, D. Schreurs, and A. Santarelli, "Wideband active load-pull by device output match compensation using a vector network analyzer," *IEEE Transactions on Microwave Theory and Techniques*, pp. 1–1, 2020.

[187] G. P. Gibiino, A. Maria Angelotti, A. Santarelli, and P. A. Traverso, "Broadband error vector magnitude characterization of a GaN power amplifier using a vector network analyzer," in *24th IMEKO TC4 International Symposium*, 2020.

[188] V. Gillet, M. Bouslama, J. Teyssier, M. Prigent, J. Nallatamby, and R. Quéré, "An unequally spaced multi-tone load–pull characterization technique for simultaneous linearity and efficiency assessment of rf power devices," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 2505–2513, 2019.

[189] L. Galatro, M. Marchetti, and M. Spirito, "60 ghz mixed signal active load-pull system for millimeter wave devices characterization," in *80th ARFTG Microwave Measurement Conference*, pp. 1–6, 2012.

[190] S. Linge and H. P. Langtangen, *Solving Nonlinear Algebraic Equations*, pp. 177–201. Springer International Publishing, 2016.

[191] D. Nopchinda, T. Eriksson, H. Zirath, and K. Buisman, "Measurement of reflection and transmission coefficients using finite impulse response least-squares estimation," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 1, pp. 222–235, 2020.

[192] M. Thorsell and K. Andersson, "Fast multiharmonic active load–pull system with waveform measurement capabilities," *IEEE Trans. Microw. Theory Techn.*, vol. 60, pp. 149–157, Jan 2012.

[193] K. Lukasik, P. Barmuta, T. Nielsen, W. Wiatr, and D. Schreurs, "Identification of multitone x-parameters under variable random phase wideband excitations," *Int. Conf. Microw., Radar and Wirel. Comm. (MIKON)*, 2020.

[194] W. Van Moer and Y. Rolain, "Best linear approximation: Revisited," in *IEEE Int. Instrum. Meas. Tech. Conf. (I2MTC)*, pp. 110–113, 2009.

[195] J. M. Horn, J. Verspecht, D. Gunyan, L. Betts, D. E. Root, and J. Eriksson, "X-parameter measurement and simulation of a gsm handset amplifier," in *Proc. Eur. Microw. Int. Circ. Conf.*, pp. 135–138, Oct 2008.

[196] R. Pintelon and J. Schoukens, *Design of Excitation Signals*, ch. 4, pp. 115–138. John Wiley and Sons, 2005.

[197] K. Łukasik, J. Cheron, G. Avolio, A. Lewandowski, D. F. Williams, W. Wiatr, and D. M. M. . Schreurs, "Uncertainty in large-signal measurements under variable load conditions," *IEEE Trans. Microw. Theory Techn.*, pp. 1–1, 2020.

[198] F. Giannini, P. Colantonio, G. Orengo, A. Serino, G. Stegmayer, M. Pirola, and G. Ghione, "Neural networks and Volterra series for time-domain power amplifier behavioral models," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 17, no. 2, pp. 160–168, 2007.

[199] M. Isaksson and D. Rönnow, "A parameter-reduced Volterra model for dynamic RF power amplifier modeling based on orthonormal basis functions," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 17, no. 6, pp. 542–551, 2007.

[200] P. N. Landin, M. Isaksson, and P. Händel, "Parameter extraction and performance evaluation method for increased performance in RF power amplifier behavioral modeling," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 20, no. 2, pp. 200–208, 2010.

[201] Y. Ma, Y. Akaiwa, Y. Yamao, and S. He, "Test bed for characterization and predistortion of power amplifiers," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 23, no. 1, pp. 74–82, 2013.

[202] F. M. Ghannouchi, O. Hammi, and M. Helaoui, *Behavioral modeling and predistortion of wideband wireless transmitters*. John Wiley & Sons, 2015.

[203] S. Yan, C. Zhang, and Q.-J. Zhang, "Recurrent neural network technique for behavioral modeling of power amplifier with memory effects," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 25, no. 4, pp. 289–298, 2015.

[204] J. Cai, R. Gonçalves, and J. C. Pedro, "A new complex envelope behavioral model for load mismatched power amplifiers," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 27, no. 6, p. e21097, 2017.

[205] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3852–3860, 2006.

[206] P. Gilabert, G. Montoro, and E. Bertran, "On the Wiener and Hammerstein models for power amplifier predistortion," in *2005 Asia-Pacific Microwave Conference Proceedings*, vol. 2, pp. 4 pp.–, 2005.

[207] D. Mirri, G. Luculano, F. Filicori, G. Pasini, G. Vannini, and G. Gabriella, "A modified Volterra series approach for nonlinear dynamic systems modeling," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, no. 8, pp. 1118–1128, 2002.

[208] E. Ngoya, C. Quindroit, and J. M. Nebus, "On the continuous-time model for nonlinear-memory modeling of RF power amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 12, pp. 3278–3292, 2009.

[209] J. Verspecht, J. Horn, L. Betts, D. Gunyan, R. Pollard, C. Gillease, and D. E. Root, "Extension of X-parameters to include long-term dynamic memory effects," in *2009 IEEE MTT-S International Microwave Symposium Digest*, pp. 741–744, IEEE MTT-S International Microwave Symposium Digest, 2009.

[210] E. Ngoya and A. Soury, "Modeling memory effects in nonlinear subsystems by dynamic Volterra series," in *Proceedings of the 2003 IEEE International Workshop on Behavioral Modeling and Simulation*, pp. 28–33, Proceedings of the 2003 IEEE International Workshop on Behavioral Modeling and Simulation, 2003.

[211] C. Quindroit, E. Ngoya, A. Bennadji, and J.-M. Nebus, "An orthogonal lookup-table decomposition for accurate IMD prediction in power amplifier with memory," in *2008 IEEE MTT-S International Microwave Symposium Digest*, pp. 1437–1440, IEEE MTT-S International Microwave Symposium Digest, 2008.

[212] A. M. Angelotti, G. P. Gibiino, and A. Santarelli, "Nonlinear dynamic modeling of RF PAs using custom vector fitting algorithm," in *2018 International Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits (INMMIC)*, pp. 1–3, 2018.

[213] A. M. Angelotti, G. P. Gibiino, and A. Santarelli, "Efficient implementation of a modified-Volterra radio-frequency power amplifier nonlinear dynamic model by global rational functions approximation," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 30, no. 11, p. e22392, 2020.

[214] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *IEEE Transactions on power delivery*, vol. 14, no. 3, pp. 1052–1061, 1999.

[215] F. Ferranti, D. Deschrijver, L. Knockaert, and T. Dhaene, "Hybrid algorithm for compact and stable macromodelling of parameterised frequency responses," *Electronics letters*, vol. 45, no. 10, pp. 493–495, 2009.

[216] F. Ferranti, L. Knockaert, and T. Dhaene, "Parameterized S-parameter based macromodeling with guaranteed passivity," *IEEE Microwave and Wireless Components Letters*, vol. 19, no. 10, pp. 608–610, 2009.

[217] F. Ferranti, Y. Rolain, L. Knockaert, and T. Dhaene, "Variance weighted vector fitting for noisy frequency responses," *IEEE Microwave and Wireless Components Letters*, vol. 20, no. 4, pp. 187–189, 2010.

[218] J. B. King and T. J. Brazil, "Time-domain simulation of passband S-parameter networks using complex baseband vector fitting," in *2017 Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC)*, pp. 1–4, Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC), 2017.

[219] A. Bennadji, A. Layec, A. Soury, A. Mallet, E. Ngoya, and R. Quere, "Modeling of a communication chain with implementation of a Volterra power amplifier model for efficient system level simulation," in *The European Conference on Wireless Technology, 2005.*, pp. 101–104, The European Conference on Wireless Technology, 2005.

[220] D. Gapillout, C. Mazière, E. Ngoya, and S. Mons, "A reliable methodology for experimental extraction of power amplifier dynamic Volterra model," in *2015 Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC)*, pp. 1–3, Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC), 2015.

[221] G. P. Gibiino, J. Couvidat, G. Avolio, D. Schreurs, and A. Santarelli, "Supply-terminal 40 MHz BW characterization of impedance-like nonlinear functions for

envelope tracking PAs," in *2016 87th ARFTG Microwave Measurement Conference (ARFTG)*, pp. 1–4, Proc. ARFTG Microwave Measurement Conference, 2016.

[222] P. A. Traverso, D. Mirri, G. Pasini, and F. Filicori, "A nonlinear dynamic S/H-ADC device model based on a modified Volterra series: Identification procedure and commercial cad tool implementation," *IEEE Transactions on Instrumentation and Measurement*, vol. 52, no. 4, pp. 1129–1135, 2003.

[223] A. Zhu, J. C. Pedro, and T. J. Brazil, "Dynamic deviation reduction-based Volterra behavioral modeling of RF power amplifiers," *IEEE Transactions on microwave theory and techniques*, vol. 54, no. 12, pp. 4323–4332, 2006.

[224] B. Gustavsen, "The vector fitting website," *MATLAB code, http://www. sintef. no/vectfit*, vol. 21, 2013.

[225] S. A. Albahrani *et al.*, "Characterization of trapping in a GaN HEMT by performing isothermal three-stage pulse measurements," in *Proc. European Microw. Int. Circ. Conf. (EuMIC)*, pp. 161–164, Oct. 2016.

[226] P. Barmuta, G. P. Gibiino, F. Ferranti, A. Lewandowski, and D. M. M. . Schreurs, "Design of experiments using centroidal Voronoi tessellation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 11, pp. 3965–3973, 2016.

[227] G. P. Gibiino, A. Santarelli, and F. Filicori, "A GaN HEMT global large-signal model including charge trapping for multibias operation," *IEEE Trans. Microw. Theory Techn.*, vol. 66, pp. 4684–4697, Nov. 2018.

[228] V. Vadalà, G. Avolio, A. Raffo, D. M.-P. Schreurs, and G. Vannini, "Nonlinear embedding and de-embedding techniques for large-signal FET measurements," *Microwave and Optical Technology Letters*, vol. 54, no. 12, pp. 2835–2838, 2012.

[229] A. Benvegnù *et al.*, "On-wafer single-pulse thermal load–pull RF characterization of trapping phenomena in AlGaN/GaN HEMTs," *IEEE Trans. Microw. Theory Techn.*, vol. 64, pp. 767–775, Mar. 2016.

[230] A. Maria Angelotti, G. P. Gibiino, T. Nielsen, J. Verspecht, and A. Santarelli, "Combined wideband active load-pull and modulation distortion characterization with avector network analyzer," in *Accepted for 97th ARFTG Microwave Measurement Conference*, 2021.

[231] G. P. Gibiino, A. M. Angelotti, A. Santarelli, F. Filicori, and P. A. Traverso, "Multitone multiharmonic scattering parameters for the characterization of nonlinear networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[232] P. Barmuta, K. Lukasik, F. Ferranti, G. P. Gibiino, A. Lewandowski, and D. Schreurs, "Load-pull measurements using centroidal Voronoi tessellation," in *2017 89th ARFTG Microwave Measurement Conference (ARFTG)*, pp. 1–4, 2017.

[233] P. Barmuta, F. Ferranti, K. Łukasik, A. Lewandowski, and D. Schreurs, "Concurrent surrogate modeling and adaptive sampling in load-pull measurements," in *2015 Integrated Nonlinear Microwave and Millimetre-wave Circuits Workshop (INMMiC)*, pp. 1–3, 2015.

[234] S. Amin, P. N. Landin, P. Händel, and D. Rönnow, "Behavioral modeling and linearization of crosstalk and memory effects in RF MIMO transmitters," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 4, pp. 810–823, 2014.

[235] P. L. Gilabert, G. Montoro, D. Vegas, N. Ruiz, and J. A. Garcia, "Digital predistorters go multidimensional: Dpd for concurrent multiband envelope tracking and outphasing power amplifiers," *IEEE Microwave Magazine*, vol. 20, no. 5, pp. 50–61, 2019.

[236] D. López–Bueno, Q. A. Pham, G. Montoro, and P. L. Gilabert, "Independent digital predistortion parameters estimation using adaptive principal component analysis," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 12, pp. 5771–5779, 2018.

[237] J. A. Becerra, M. J. Madero-Ayora, J. Reina-Tosina, C. Crespo-Cadenas, J. García-Frías, and G. Arce, "A doubly orthogonal matching pursuit algorithm for sparse predistortion of power amplifiers," *IEEE Microwave and Wireless Components Letters*, vol. 28, no. 8, pp. 726–728, 2018.

[238] P. Jaraut, M. Rawat, and F. M. Ghannouchi, "Composite neural network digital predistortion model for joint mitigation of crosstalk, I/Q imbalance, nonlinearity in MIMO transmitters," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 11, pp. 5011–5020, 2018.

[239] M. B. Yelten, T. Zhu, S. Koziel, P. D. Franzon, and M. B. Steer, "Demystifying surrogate modeling for circuits and systems," *IEEE Circuits and Systems Magazine*, vol. 12, no. 1, pp. 45–63, 2012.

[240] M. Mengozzi, G. P. Gibiino, A. M. Angelotti, C. Florian, and A. Santarelli, "Supply-modulated PA performance enhancement by joint optimization of rf input and supply control," in *2020 IEEE Asia-Pacific Microwave Conference (APMC)*, pp. 585–587, 2020.