Wideband Active Load-Pull by Device Output Match Compensation Using a Vector Network Analyzer

Wideband Active Load-Pull by Device Output Match Compensation using a Vector Network Analyzer

Alberto M. Angelotti, *Graduate Student Member, IEEE,* Gian Piero Gibiino, *Member, IEEE,*
Troels S. Nielsen, *Member, IEEE,* Dominique Schreurs, *Fellow, IEEE,* and Alberto Santarelli, *Member, IEEE*

***Abstract*—This work investigates wideband active load-pull (WALP) on microwave electron devices with a standard vector network analyzer (VNA), solely using calibrated frequency-domain relative measurements. Differently from the methods requiring full time-domain waveform acquisition capabilities, this approach removes any instantaneous bandwidth (BW) requirement, allowing to target modulated signals with arbitrarily large BWs. A VNA-based measurement setup is presented, and a novel mathematical framework is developed in order to enable the synthesis of an arbitrary wideband load profile using suitable numerical algorithms. A linear pre-compensation of the LS output match of the device, here implemented by different methods, is shown to significantly improve the speed and the stability of the basic secants' iterative method used to reach the target. Performance results, investigating the convergence properties of the approach and the effects of the setup dynamic range, are reported for a Gallium Nitride (GaN) high-electron-mobility-transistor (HEMT) for up to 120-MHz output BW in the sub-6 GHz range.**

***Index Terms*—Active load-pull, modulated signals, power amplifiers, vector network analyzer, device characterization.**

I. Introduction

THE ever-growing demand for high data rates in modern mobile communication standards, e.g., 5G, requires the operation of power amplifiers (PA) in radio frequency (RF) transmitters across extremely wide bandwidths (BWs), i.e., tens or hundreds of MHz. In this respect, novel characterization techniques are required to evaluate the behaviour of electron devices used in PA circuits under operating conditions that closely approximate typical telecom applications. Wideband active load-pull (WALP) has been proposed as a promising technique to characterize microwave transistors and to optimize PA performance using excitations that mimic typical communication standards [1]–[4]. The method allows to set, for a given device-under-test (DUT), a user-prescribed load reflection coefficient profile across a wide modulation BW at fundamental and harmonic frequencies. The approach, at the cost of increased complexity, overcomes the main disadvantages of traditional passive load-pull. WALP can provide full coverage of the Smith Chart compensating for insertion loss, and can extend the narrow relative BW due to delay and resonant behavior of passive tuners, while at the same time, featuring generally faster sweep times. It also allows to replicate the active load-modulation conditions often realized in application-like scenarios, such as in complex PA architectures [5], [6], as well as providing complete datasets for PA modeling [7]. Several commercial solutions for WALP are available [2], [4], which make use of time-domain waveform demodulation capabilities exploiting broadband acquisition hardware (HW) [1]–[4].

The main novelty of this work is to enable WALP under modulated excitations by leveraging on standard vector network analyzer (VNA) measurements without using any time-domain waveform reconstruction capabilities. 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 demodulation of large instantaneous BWs, enabling WALP on the much wider excitation BWs required by modern communication standards. Thanks to these novel capabilities, several VNA-based advanced PA measurements, such as wideband modulation distortion [8]–[10], can be performed directly at transistor level. Indeed, the VNA-WALP proposed in this work allows to emulate the output termination conditions seen in the final application, without needing any detailed knowledge of the modulated loadline. This permits to efficiently characterize several key performance metrics at an early stage of design, before the physical realization of the actual PA circuit.

On the other hand, this work does not aim at depicting any particular PA design strategy, neither it targets the identification of a single specific optimal termination impedance. 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 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.

With respect to the preliminary article in [9], this extended work further clarifies the proposed theoretical framework and validates the capabilities of the VNA-WALP method by reporting novel measurement results. The performance of the secants' method, used to set the required load profile, is

A. M. Angelotti, G. P. Gibiino, and A. Santarelli are with the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi," University of Bologna, 40136 Bologna, Italy (e-mail: alberto.angelotti@unibo.it).

T. S. Nielsen is with Keysight Technologies, 1400 Fountaingrove PKY, Santa Rosa CA-95403, USA.

D. Schreurs is with the TELEMIC Division of the Department of Electrical Engineering (ESAT), KU Leuven, Leuven 3001, Belgium.

compared to other known numerical algorithms. This basic method is further extended by using a large-signal (LS) output match compensation procedure, whose improvements in performance are experimentally validated using realistic wideband excitations. Moreover, the impact of passive pre-matching conditions is examined and behavioral modeling aspects of the LS output match of the DUT are discussed in detail. The impact of the measurements' dynamic range on the stability and accuracy of the VNA-WALP method is theoretically and experimentally studied. This analysis is then employed to explore the possibility of performing VNA-WALP on wider modulation BWs that include spectral re-growth components.

The article is organized as follows. The proposed framework is introduced in Sec. II, while the measurement setup is presented in Sec. III. A comparison between different iterative algorithms used to set the target load profile i s presented in Sec. IV. Section V introduces a novel approach, using a reduced number of measurements, to improve the speed and stability of the algorithm by compensating for the LS output match of the device. Section VI compares the performance of this approach to the basic secants' algorithms introduced in Sec. IV. Section VII analyzes the achievable accuracy in the proposed set-up, and extends the method to out-of-band WALP. Finally, conclusions are drawn in Sec. VIII.

II. Problem Statement

Let us consider the case in which the DUT is excited at the input by a user-defined p eriodic b and-pass s ignal $a_{S1}(t)$ 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 [11]. Among the many different techniques available in literature [8], [11]–[13], flat-amplitude r andom-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 periodic-in-same-period-out (PISPO) behaviour [14], 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 [2], [3], this work will only consider load-pull for a certain BW around a given RF carrier. Overall, the frequency domain measurements will include both the input-excited tones and the spectral regrowth such as third or higher order intermodulation (IM) distortion products. The reflection c oefficients se en by th e DU T at ba seband 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.

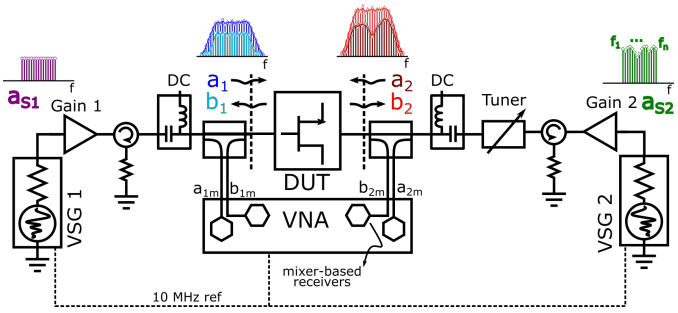


Fig. 1. Block diagram of the WALP setup.

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. 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 [1]–[4], [12]. 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 $A_2(f) = B_2(f) \cdot \Gamma_T(f)$.

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 $b_2(t)$. Instead, just calibrated amplitude spectra and same-frequency ratioed measurements are exploited in order to set the required target. Thus, 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:

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

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

$\underline{A}_2$ and B_2 waves. By defining the frequency-domain vector $A_{S2} \stackrel{def}{=} [A_{S2}(f_1) \dots A_{S2}(f_N)]^t$, (1) can be recast as:

$$E_n(\overline{A}_{S2}) \stackrel{def}{=} G_n(\overline{A}_{s2}) - \Gamma_T(f_n) = 0, \qquad n = 1 \dots N; \tag{2}$$

where E_n is the reflection coefficient error at frequency f_n. Equation (1) can be turned into a system of N complex non-linear equations in N complex variables:

$$\overline{E}(\overline{A}_{S2}) = \left[E_1(\overline{A}_{S2}) \dots E_N(\overline{A}_{S2})\right]^T = \overline{0}; \tag{3}$$

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 of the error function) to (3). This is a classical problem in mathematics, with several numerical approaches reported in literature [15]. 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 [16] could 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. III) across all the frequencies at which load-pull is performed. 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. IV).

III. Measurement Setup

The block diagram and picture of the WALP setup used in this work are shown in Figs. 1 and 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

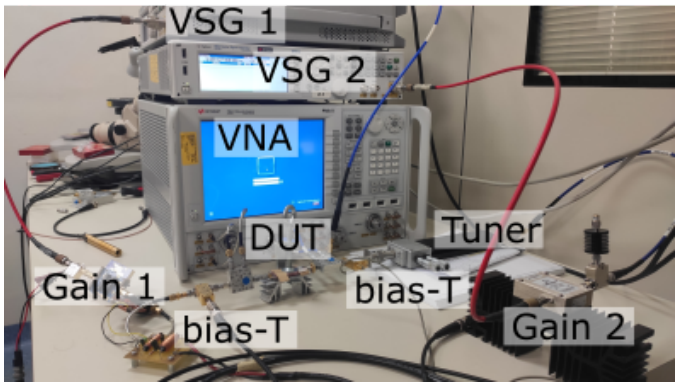


Fig. 2. A picture of the proposed VNA-WALP measurement setup.

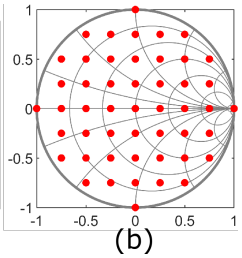
decoupled from the DUT using circulators. These amplifiers should display highly-linear operation across the whole power range of interest in order to avoid unwanted nonlinear distortion between the digital signals stored in VSG memory and the actual waveforms at the DUT reference planes. While minor nonlinear components can be automatically compensated during the iterative load-setting procedure (Sec. IV), 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 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. II, 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 [2], [4] 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 [8], [12], [17]. 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

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

(a)



(b)

Fig. 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.

TABLE I
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

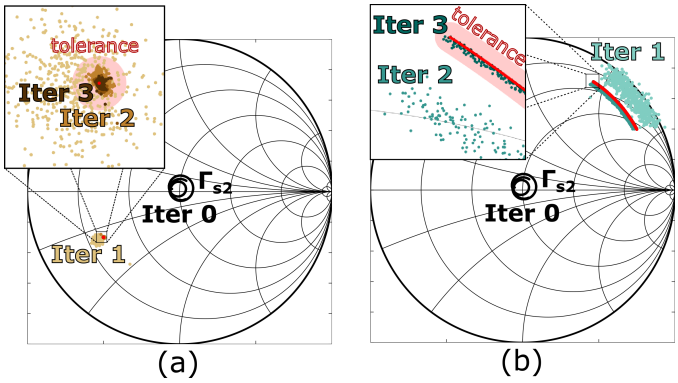
control and for processing WALP algorithms are implemented in MATLAB®, running on an external LAN-networked computer.

IV. WALP METHODS' COMPARISON

The performance of the VNA-WALP depends on the adopted algorithm to solve (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. A detailed description of each method, together with the implications on the WALP case, can be found in Appendix A.

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 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



(a) (b)

Fig. 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.

(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. 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. I 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 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 [8], [11], [13], [16], [17].

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 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. 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

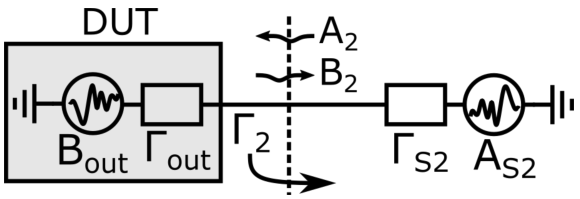


Fig. 5. Linear model for the DUT and output injection source at the load reference plane.

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 (3).

V. Device Output Match Compensation

A. Experimental procedure

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 (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. IV), 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. IV, 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. 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:

$$E_n = \frac{\Gamma_{S2}(f_n) + \eta_n}{1 + \Gamma_{out}(f_n)\eta_n} - \Gamma_T(f_n); \quad \eta_n \stackrel{def}{=} \frac{A_{S2}(f_n)}{B_{out}(f_n)}; \qquad (4)$$

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. The derivation of (4) is reported in Appendix B. Equation (4) shows that, even in the case of a perfectly linear DUT, for which diagonality of $\mathbf{J}$ 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) 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) 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) 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 (LSOP) 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:

$$\Gamma_{S2}(f) = \left.\frac{A_2(f)}{B_2(f)}\right|_{@inEXC}. \qquad (5)$$

   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:

$$\Gamma_{out}(f) = \left.\frac{B_2(f)}{A_2(f)}\right|_{@outEXC}. \qquad (6)$$

   The behavioral meaning of this LS output match parameter will be discussed in Sec. V-C.
3) The measured Γ_{out} from (6) is smoothly interpolated across frequency, in order to provide estimation of its value at the input excited (i.e., non-offset) frequencies.

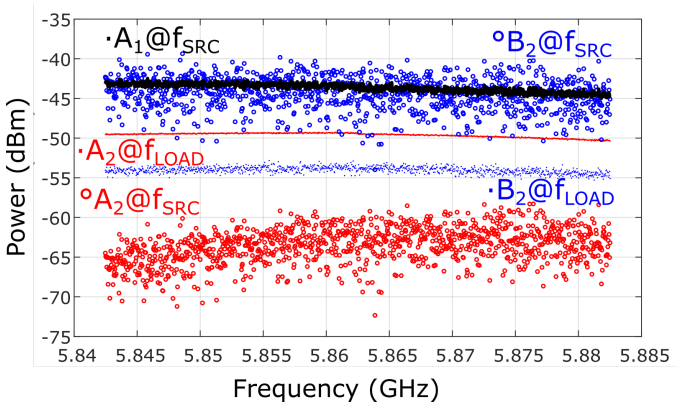


Fig. 6. Interleaved excitation strategy for the estimation of Γ_{S2} and Γ_{out} on 40-MHz excitation BW for a GaN HEMT. A_1 (black), A_2 (red) and B_2 (blue) at source-side (circles) and load-side (dots) excited frequencies are reported.

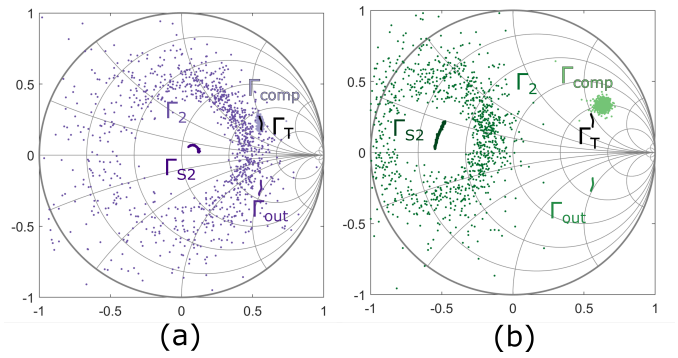


Fig. 7. Reflection coefficients measured at different steps of the output match compensation procedure for two different output source matches. (a) Γ_{S2} obtained by the output PA and its circulator directly connected to the DUT. (b). Γ_{S2} synthesized using a passive tuner between the output source and DUT.

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 (5), similarly to the secants' method starting point. This step allows to estimate the current injection ratio using (4) as:

$$\eta(f) = \frac{\Gamma_2(f_n) - \Gamma_{S2}(f_n)}{1 - \Gamma_{out}(f_n)\Gamma_2(f_n)}, \tag{7}$$

where the measurement of the reflection coefficient is used:

$$\Gamma_2(f) = \left.\frac{A_2(f)}{B_2(f)}\right|_{@\,in/outEXC}. \tag{8}$$

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.

$$\eta_T(f) = \frac{\Gamma_T(f) - \Gamma_2(f)}{1 - \Gamma_{out}(f)\Gamma_T(f)} + \eta(f)\frac{1 - \Gamma_{out}(f)\Gamma_2(f)}{1 - \Gamma_{out}(f)\Gamma_T(f)}. \tag{9}$$

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:

$$A_{S2}^{comp}(f_n) = A_{S2}^{0}(f_n)\frac{\eta_T(f_n)}{\eta(f_n)}, \tag{10}$$

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

B. 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 the Introduction, finding such an optimum by sweeping all impedance combinations for a very 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 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. 7 for two different source matches. In the first case (Fig. 7a), the output PA with its circulator are directly connected to the DUT. In the second one (Fig. 7b), a passive tuner is added between the source and DUT (see Figs. 1 and 2) in order to emulate a load pre-match condition, such as the one found in hybrid passive-active load-pull setups.

These experimental tests show that the procedure is able to provide, in both cases, 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, together with the ones reported in [9] for 80-MHz excitation BW and $N = 237$ tones, prove that the linear compensation procedure is able to correct for the effect of the DUT mismatch using very few measurements, and that it is directly applicable to hybrid load-pull scenarios.

C. 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 (6), 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[1] framework [18], [19], 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 [18]. Instead, the best-linear approximation (BLA) framework [20] and the hot-S_{22} approach [21] 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. V-A, uses a further modification of the previous methods, exploiting a SS tickler multitone [22] 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 [23] 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 [19]. 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 [8], the measured

[1] X-parameters is a registered trademark of Keysight Technologies.

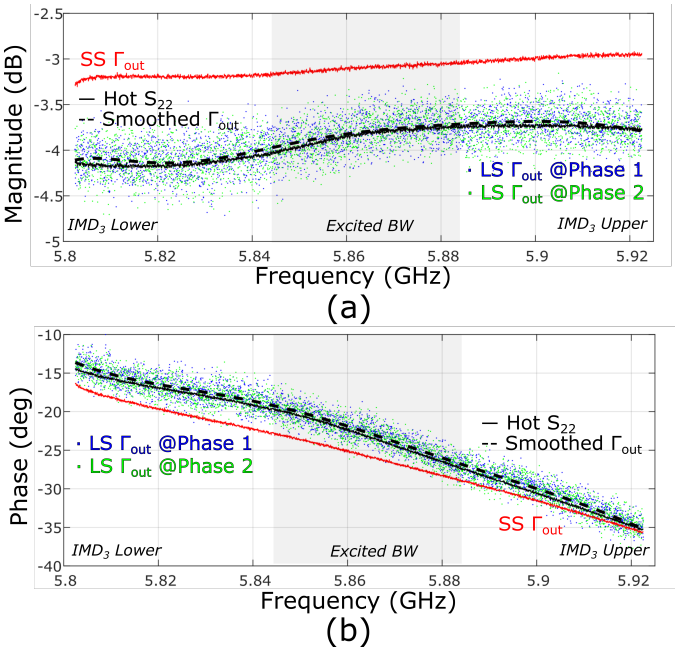


Fig. 8. Comparison of the magnitude (a) and (b) phase of different definitions of the LS output match. SS Γ_{out} (red), hot-S_{22}-like Γ_{out} (dashed black) are compared with a smoothed version (solid black) of the multitone LS Γ_{out} measurement, as used in Sec. V-A. Two multitone LS Γ_{out} measurements for two different (green and blue) phase realizations are reported. Grey shading identifies the input-excited 40-MHz BW and third-order IM BW.

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-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. 8 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 [20], allows to estimate Γ_{out} at the IM3 frequencies.

The multitone LS Γ_{out} presents LSOP-dependendent, *noise-like* nonlinear stochastic distortions [14], [22] due to the overlap of direct and IM components. The value of the LS multitone 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 [8]. In this way, the smoothed profile (referred to as

smooth Γ_{out} in Fig. 8) 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. V-A. 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. V-B and the ones shown in [9], 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 [8], 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 [14]) 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 when arbitrarily wider BWs are considered.

VI. Hybrid Method Performance

As pointed out in the previous sections and in [9], 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. V-A. 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.

In order to compare the performance of the standard secants' method with the proposed hybrid one, the GaN HEMT under test is excited in the same conditions reported in Sec. V-B. The two algorithms are run for 10 iterations without imposing any other stopping criterion in order to evaluate the intrinsic performance of each method. The adopted figures-of-merit to run the comparison are the frequency-by-frequency error and the RMS error across frequencies, defined as:

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |\Gamma_2(f_n) - \Gamma_T(f_n)|^2}. \qquad (11)$$

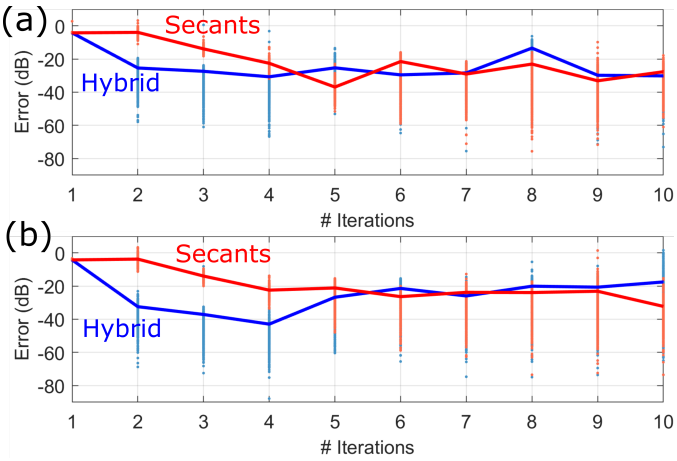


Fig. 9. Comparison of the secants' (red) and hybrid (blue) method in the WALP of a 1111-tone 40-MHz wide multitone at 5.625 GHz. The two methods are both run for 10 iterations, and the frequency-by-frequency error (dots) and the RMS error (solid line) are reported.

Figures 9a-b report the comparison of the error for 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 therefore 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. VII). 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. 9, 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. 9, 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. IV 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.

VII. VNA-WALP OPERATION AT LOW SIGNAL POWER

A. 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 [2], [4], [24]. Indeed, as described in Sec. II, 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. 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 [16]. 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 dynamic range of modern VNAs, which can be further increased by using narrower Intermediate Frequency (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 (see Fig. 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. 10. As expected, the empirical uncertainty in the Γ_2 due to measurement noise decreases with

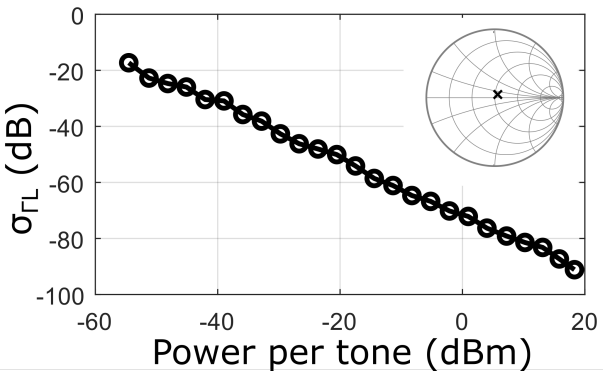


Fig. 10. Measurement noise standard deviation in function of power per tone at 5.625 GHz. The load reflection, in the Smith Chart inset, is the one seen by the DUT when active injection is turned off.

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 [25]. Figure 10 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 (see Sec. IV). 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. 10) 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.

B. 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 [8], [14] 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, for the GaN HEMT under test, in Fig. 11 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

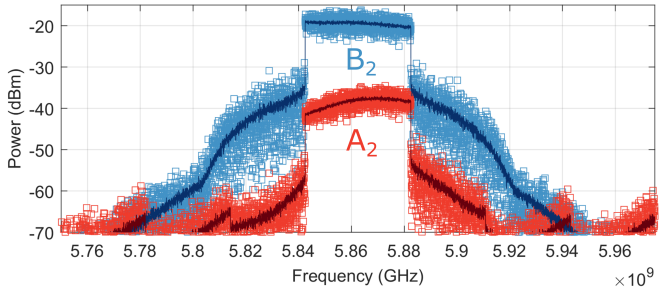


Fig. 11. PSD across frequency for the A_2 (red) and B_2 (blue) waves for the GaN HEMT under test when excited by a $N = 1111$ flat-amplitude random-phase multitone signals. Darker continuous lines represent a statistical average of the PSD of the stochastic processes, while lighter squares show the particular realization.

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 [8] or other gaussian signals [14], [16]. 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 [14], [17], 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 [8] is to fix a number N of tones, typically in the thousands range, that will allow to match the ccdf of the application 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. 10. In some applications, however, the tone spacing might influence the long-term memory behaviour of the DUT [17] 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.

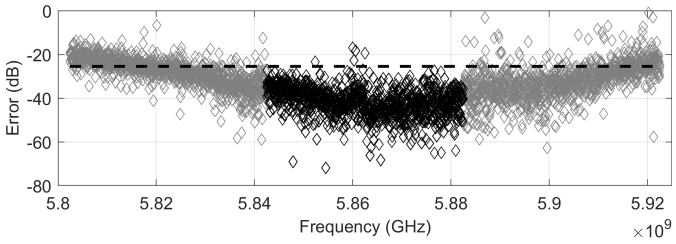


Fig. 12. Error across frequency for a 40-MHZ input random-phase excitation and a 120 MHz VNA-WALP across the third order output BW. Input excited BW (black), third-order IM BW (grey) and RMS error across frequency (dashed line) are reported.

C. 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. 12. 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. 8. 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.

VIII. Conclusion

In this work, 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 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 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.

APPENDIX A
ITERATIVE METHODS FOR VNA-BASED WALP

1) 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 [3], [26]. 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:

$$\overline{A}_{S2}^{(r)} = \overline{A}_{S2}^{(r-1)} - \mathbf{J}_{h,k}\left(\overline{A}_{S2}^{(r-1)}\right)\overline{E}\left(\overline{A}_{S2}^{(r)}\right). \tag{12}$$

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.

2) Nonlinear least-squares: the system in (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.

3) 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.

4) Secants' Method: in the case of a perfectly linear DUT, any injected tone would cause a response just at the exact same frequency, without any regrowth or interaction among the different tones due to IM distortion. 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 [15] 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 in which no injection ($\overline{A}_{S2} = 0$) is applied. For the second one, we select a SS 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.

APPENDIX B
DERIVATION OF Γ_T ERROR IN (4)

Referring to the schematic in Fig. 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):

$$B_2 = B_{out} + \Gamma_{out} A_2. \tag{13}$$

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

$$A_2 = A_{S2} + \Gamma_{S2} B_2. \tag{14}$$

By solving the two equations (13) and (14) for the two unknown waves A_2 and B_2 results in

$$A_2 = \frac{A_{s2} + \Gamma_{s2} B_{out}}{1 - \Gamma_{s2}\Gamma_{out}}; \quad B_2 = \frac{B_{out} + \Gamma_{out} A_{s2}}{1 - \Gamma_{s2}\Gamma_{out}}. \tag{15}$$

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

$$\begin{aligned} \Gamma_2 \overset{\text{def}}{=} \frac{A_2}{B_2} &= \frac{A_{s2} + \Gamma_{s2} B_{out}}{B_{out} + \Gamma_{out} A_{s2}} \\ &= \frac{\frac{A_{s2}}{B_{out}} + \Gamma_{s2}}{1 + \Gamma_{out}\frac{A_{s2}}{B_{out}}} = \frac{\eta + \Gamma_{s2}}{1 + \Gamma_{out}\eta} \end{aligned} \tag{16}$$

where η is defined as the ratio between the load injection source and the DUT output-equivalent source $\eta \overset{\text{def}}{=} \frac{A_{s2}}{B_{out}}$.

ACKNOWLEDGMENT

This work is supported in part by Keysight Technologies, Santa Rosa (US).

REFERENCES

[1] 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, Oct 2004, pp. 685–688.

[2] 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, no. 12, pp. 2979–2988, Dec 2008.

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

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

[5] 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.

[6] 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.*, June 2017, pp. 947–949.

[7] 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 Trans. Microw. Theory Techn.*, vol. 66, no. 2, pp. 831–844, 2017.

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

[9] A. M. Angelotti, G. 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 *IEEE MTT-S Int. Microw. Symp. Dig.*, June 2020.

[10] G. P. Gibiino, A. Angelotti, A.Santarelli, and P.A.Traverso, "Vna-based broadband evm measurement of an rf nonlinear pa under load mismatch conditions," in *IMEKO-TC4 International Symposium*, Sept. 2020.

[11] 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, no. 3, pp. 122–138, June 2008.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] A. M. Angelotti, G. Gibiino, C. Florian, and A.Santarelli, "Broadband error vector magnitude characterization of a gan power amplifier using a vector network analyzer," in *IEEE MTT-S Int. Microw. Symp. Dig.*, June 2020.

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

[19] 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.

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

[21] 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.*, Oct 2008, pp. 135–138.

[22] 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, no. 12, pp. 4133–4146, Dec 2018.

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

[24] 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.

[25] 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.

[26] M. Thorsell and K. Andersson, "Fast multiharmonic active load–pull system with waveform measurement capabilities," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 1, pp. 149–157, Jan 2012.

Alberto Maria Angelotti (Graduate Student Member, IEEE) received his BSc and MSc degrees in Electronics from the University of Bologna, Bologna, Italy in 2014 and 2017, respectively. Since 2017, he has been with the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi," University of Bologna, where he is currently pursuing a PhD degree. His research interests include microwave instrumentation, nonlinear measurements, and behavioral modeling of power amplifiers.

Gian Piero Gibiino Gian Piero Gibiino (Member, IEEE) received the dual Ph.D. degree from the University of Bologna, Bologna, Italy, and KU Leuven, Leuven, Belgium, in 2016.,Since 2012, he has been with the Department of Electrical Engineering "Guglielmo Marconi"-DEI, University of Bologna, where he is currently a Post-Doctoral Research Fellow. His research interests include RF electron devices and power amplifiers nonlinear modeling, electronic instrumentation, and microwave measurements.,Dr. Gibiino is a member of the Italian Association of Electrical and Electronic Measurements (GMEE).

Troels Nielsen received the master's and Ph.D. degrees in electrical and electronic engineering from Aalborg University, Aalborg, Denmark, in 2002 and 2006, respectively. He is currently a Research Scientist with the Measurement Research Laboratory, Keysight Technologies, Santa Rosa, CA, USA. From 2005 to 2009, he was a Senior RF Design Engineer with the Corporate Research and Development Modeling Group, RFMD, Greensboro, NC, USA. He is involved in research and development of large-signal nonlinear models for III–V technology power amplifiers and large-signal measurements for nonlinear model development and validation. He has authored a dozen technical papers, articles, and book contributions within the fields of RF/microwave IC design, characterization, and modeling. His current research interests include techniques for system-level modeling, nonlinear system identification techniques, large-signal nonlinear measurements, and power amplifier linearization techniques.

Dominique Schreurs (Fellow, IEEE) received the M.Sc. degree in electronic engineering and the Ph.D. degree from the University of Leuven (KU Leuven), Leuven, Belgium, in 1992 and 1997, respectively. She has been a Visiting Scientist with Agilent Technologies, Santa Rosa, CA, USA, ETH Zürich,Zürich, Switzerland, and the National Institute of Standards and Technology, Boulder, CO, USA. She is currently a Full Professor with KU Leuven, where she is also the Chair of the Leuven ICT. Her current research interests include the microwave and millimeter-wave characterization and modeling of transistors, nonlinear circuits, and bioliquids, and system design for wireless communications and biomedical applications. Prof. Schreurs served as the President of the IEEE Microwave Theory and Techniques Society from 2018 to 2019. She was an IEEE MTT-S Distinguished Microwave Lecturer. She has also served as the General Chair for the Spring Automatic RF Techniques Group (ARFTG) conferences in 2007,2012, and 2018, and the President of the ARFTG organization from 2018 to2019. She currently serves as the TPC Chair for the European Microwave Conference and also the Conference Co-Chair for the IEEE International Microwave Biomedical Conference. She was the Editor-in-Chief of the IEEE Transactions on Microwave Theory and Techniques.

Alberto Santarelli Alberto Santarelli received the Laurea degree (cum laude) in electronic engineering in 1991 and the Ph.D. in Electronics and Computer Science from the University of Bologna, Italy in 1996. He was a Research Assistant from 1996 to 2001 with the Research Centre for Computer Science and Communication Systems of the Italian National Research Council (IEIIT-CNR) in Bologna. In 2001, he joined the Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, where he currently is an Associate Professor. During his academic career he has been Lecturer of High-frequency Electronic Circuits, Applied Electronics and Power Electronics. His main research interests are related to nonlinear characterization and modeling of electron devices and to nonlinear circuit design. Prof. Santarelli is a member of the European Microwave Association (EuMA).