2022 IEEE Latin America Electron Devices Conference (LAEDC). Puebla, México, July 4-6, 2022

Effect of probe coupling on MOSFET series resistance extraction up to 110 GHz

L. Nyssens[1], M. Rack[1], D. Lederer[1], J.-P. Raskin[1]

[1]Université catholique de Louvain, Louvain-la-Neuve, Belgium, E-mail: lucas.nyssens@uclouvain.be

2022 IEEE Latin American Electron Devices Conference (LAEDC) | 978-1-6654-9767-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/LAEDC54796.2022.9908195

***Abstract*—The measurement of series extrinsic resistances of MOSFETs is not straightforward and the gate resistance (R_g) in particular is very sensitive to noise measurement and measurement inaccuracies. They are critical elements that need an accurate estimation for proper FET modeling and RF figures of merit assessment, as they limit the extrinsic cutoff and maximum oscillation frequencies (f_t, f_{max}) in deeply scaled CMOS technologies. This work compares the extrinsic resistances extraction with different off-wafer and on-wafer calibration and de-embedding methods to provide an insight on the appropriate procedure for accurate correction. Then, measurements obtained with three different probe technologies are compared. A resonance-like signature specific to each probe technology is observed, caused by unwanted coupling between the probe and the on-chip neighbor environment of the test structures. This coupling is not well corrected by the calibration and de-embedding procedure and is reflected back on the corrected measurements, mainly on the f_{max} and R_g curves. Overall, the best probe technology for DC-110 GHz measurements among the probes tested in this work is identified to be Picoprobe, featuring a gate resistance extraction with less than ±3% variation from 10 to 60 GHz.**

Keywords—source and drain resistances, gate resistance, MOSFET, CMOS, on-wafer calibration, de-embedding, millimeter-wave measurements, DC-110 GHz measurements

I. Introduction

Continuous downscaling of CMOS technology and extrinsic parasitics optimization has improved the RF figures of merit (FoMs) of Si technology. Nowadays, nMOSFETs featuring cutoff frequency (f_t) and maximum oscillation frequency (f_{max}) above 300 GHz are available in several Si technologies [1]. Such high RF FoMs enabling high-end millimeter-wave (mm-wave) operation, coupled to low-cost integration of digital and analog circuitry has made Si technology a serious contender of III-V technologies for mm-wave applications.

To benefit from these improvements for actual circuit design, accurate characterization of these transistors is crucial to validate the associated compact model. Among the two aforementioned RF FoMs, f_{max} in particular is extremely difficult to determine for different reasons: measurement noise affecting small quantities that need to be measured and mostly measurement inaccuracy due to calibration algorithm and measurement environment [1], [2]. The extrinsic gate resistance (R_g) is also very sensitive to measurement noise [3] and measurement environment in the same way as f_{max}. Whereas the source and drain extrinsic resistances (R_s and R_d, respectively) are less sensitive parameters in a similar fashion as f_t. Series resistances (R_s, R_d, and R_g) are critical MOSFET parameters at RF and mm-wave frequencies as they limit downscaling improvement in f_t (R_s and R_d) f_{max} and minimum noise figure (mainly R_g). Overall, high-frequency measurement of transistors on Si technology is still challenging, even below 110 GHz when analyzing sensitive parameters such as f_{max} and R_g [1], [2], [4].

In this paper we investigate the effect of de-embedding strategy and probe coupling on the extraction of MOSFET (from the 22FDX® technology) extrinsic series resistances, critical parameters for accurate transistor modeling, including the really sensitive R_g parameter. The paper is organized as follows. Section II briefly describes the on-wafer calibration kit along with the transistor and its de-embedding structures for high-frequency characterization. Next, we analyze the effect of calibration and de-embedding strategies on the series resistances extraction. Then, we compare the extractions made from measurements using different RF probes and correlate the R_g extraction with the f_{max} curves.

II. Description of Measured Structures

A multi-line Thru-Reflect-Line (mTRL) calibration kit (calkit) is designed along with the DUT(s) for in-situ accurate extraction. The backend-of-line contains 10 metal layers (numbered from 1 to 10 from bottom to top) and an additional AluCap layer. The transmission line (TL) signal is designed in the thick M10 and is 9 µm wide. The ground plane is made of several thin metal layers and one thick metal layer for a combined thickness of >1 µm (M1-M8). Ground sidewalls are also present at a distance of 12.7 µm from the signal line in order to comply to the minimum density design rules without needing to add any dummy fills. Due to the relatively short distance to the signal line, the transmission line is effectively implemented as a grounded coplanar waveguide (G-CPW) line. A sketch of the G-CPW cross-section is shown in Fig. 1. The dimensions are selected to achieve a 50 Ohm characteristic impedance. Lines of multiple lengths (67 µm, 207, 625, 1648 µm) are fabricated, along with dedicated Open and Short to enable wideband mTRL calibration [5]. The Open (Short) shares the same footprint and probe to probe distance as the Thru to ease the measurements, thus with an 18 µm long opening (shorted section).

The RF signal pad in AluCap is 60 µm long, 40 µm wide and 15 µm away from the ground pads. A ground shield beneath the signal pad is designed to prevent the electric field to enter in the lossy Si substrate by stacking only the 5 thin bottommost metal layers (M1-M5) for a reduced pad capacitance. The Open structure measurement yields a 15 fF shunt capacitance for the Open structure. A sketch of the Open is given in Fig. 2(a).

An nMOSFET is used as DUT in this work. It is implemented in a separate structure with the same pads and accesses as the in-situ calkit, with the same probe to probe dis-

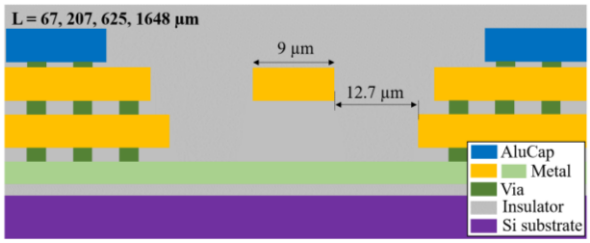

Fig. 1. Sketch of on-wafer calkit transmission line, with dimensions (not to scale).

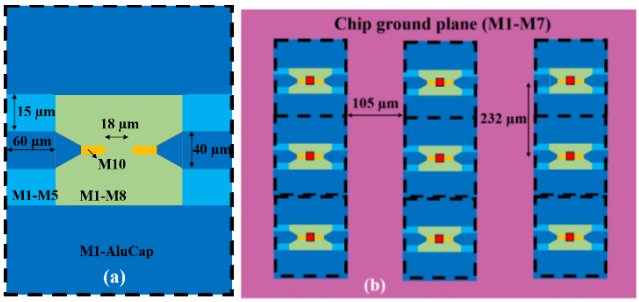

Fig. 2. (a) Sketch of on-wafer calkit Open. (b) Sketch of floor plan.

tance) as the Thru, Open and Short structures. It lies inside an opening in the TL's ground plane, thus requiring to shift the reference plane to 9 µm towards the probes after the in-situ calibration. In addition, Short and Open structures with all the FET access parasitics down to M1 are included (called Short-M1 and Open-M1) for a complete extraction down to M1.

The chip floor management is designed to present similar first neighboring structures on the left and right side of each measured structure: in-situ calkit elements, FET and its Open-M1 and Short-M1. The top and bottom structures share the same ground plane, which saves some area with a limited impact on probe coupling, since their coupling is stronger in the horizontal (x) direction [6]. The structures are separated by 105 µm in the x-direction and their ground planes are tied together by a uniform ground plane using 7 thin bottommost metal layers (M1-M7), as proposed in [7] and [8] in order to present a neighbor environment that is as close as possible for all structures (cf. Fig. 2(b)) and to eliminate the possibility of slot modes between grounds of adjacent structures.

III. Series Resistances Extraction

In this paper, we compare the measurement accuracy by using different de-embedding strategies and RF probes. To assess the validity of the corrected measurements, we focus on the series resistance extraction of a transistor. They are critical parameters to model the FET behavior and its gate resistance in particular is very sensitive and hard to extract. Bracale's method has been proved to be a robust method to extract the series extrinsic resistances in case of measurement noise [3]. It is based on the measurement of the FET in cold and strong inversion regimes (V_{ds} = 0 V, V_{gs} >> V_{th}). The measurements under such conditions (port 1 connected to the gate, port 2 to the drain) yield the following expressions:

$$Re(Z_{11} - Z_{12}) = R_g - \frac{1}{4g_d}, \quad (1)$$

$$Re(Z_{22}) = R_{sd} + \frac{1}{2g_d}, \quad (2)$$

where R_g is the extrinsic gate resistance, R_{sd} is the sum of the series source and drain resistances, g_d is the output conductance. g_d is proportional to the overdrive voltage (V_{gs}-V_{th}), such that its contribution to the above expressions becomes smaller with larger gate voltage. In Bracale's method, (1) and (2) are evalutated for several V_{gs} biases and a linear regression from their values is used to get rid of the $1/g_d$ term. The goal being to compare measurement and extraction accuracy, we only present (1) and (2) versus frequency at the highest possible V_{gs} bias of the technology, instead of extracting the actual series resistances values (R_g and R_{sd}). Nevertheless, (1) and (2) are expected to be constant with frequency (above a few GHz for (1)). Indeed complete expressions for (1) and (2) in terms of the FET small-signal equivalent circuit parameters can be found in [9]. Any frequency-dependent deviation is attributed to inaccuracy in measurement extraction.

IV. De-Embedding Strategy

The effect of de-embedding strategy is studied in details in this section. The Open-Short de-embedding is the classical de-embedding used in most cases. However, its accuracy is limited in frequency due to the distributed nature of the accesses. For high-end mm-wave measurements, the best calibration-de-embedding strategy consists of performing an on-wafer calibration with a custom on-wafer calkit to move the reference plane as close as possible to the device, followed by a classic Open-Short de-embedding [10], [11]. To keep it concise in this work, we study the effect of three different strategies: (i) 2-tier calibration with Open-Short de-embedding that serves as reference, (ii) off-wafer calibration followed by an Open-Short de-embedding, (iii) off-wafer calibration followed by a more complex 5-step de-embedding, similar to the one proposed in [6].

(i) The 2-tier calibration consists of performing a first-tier calibration on an ISS calkit (Line-Reflect-Reflect-Match in this case) to move the reference plane to the probe tips, followed by a second-tier calibration performed with a calkit embedded on the same wafer as the DUT(s). The on-wafer calibration structures has to present a neighboring environment as close as possible as DUT's close environment to include probe coupling to substrate and neighbor structures [7]. The second-tier calibration moves the reference plane at the DUT's vicinity, effectively removing the pads and some access parasitics. Then, an Open-M1-Short-M1 de-embedding is applied to remove additional parasitics down to the extrinsic accesses at M1.

The latter correction procedure is compared to an off-wafer LRRM calibration followed by two different de-embedding procedures.

(ii) In the first one, a classic Open-M1-Short-M1 de-embedding is applied, which is the usual method applied in industries.

(iii) In the last case, the off-wafer calibration is followed by a 5-step de-embedding. The complete 6-step de-embedding technique described in [6] and [11] could not be applied here due to missing de-embedding structures. Instead, the 5-step de-embedding used here consists of the following sequence (more information can be found in [6], [11]:

1) Probe-SHORT: short at probe tips (reduces calibration residuals);
2) OPEN-Top: same structure used for on-wafer (m)TRL calibration, sketch given in Fig. 2(a) (correct parallel impedance of pad and part of the access);

3) SHORT-Top: same structure used for on-wafer (m)TRL calibration, (correct series impedance of pad and part of the access);

4) OPEN-M1 (correct parallel impedance of the metal/via access above the transistor);

5) SHORT-M1 (correct series impedance of the metal/via access above the transistor);

Fig. 3 shows expressions (1) and (2) extracted with the 3 different calibration/de-embedding strategies. The value of $Re(Z_{11}-Z_{12})$ below 10 GHz is noisy and should not be considered in this work because the gate capacitance (in series with the gate resistance) dominates the Z_{11} behavior.

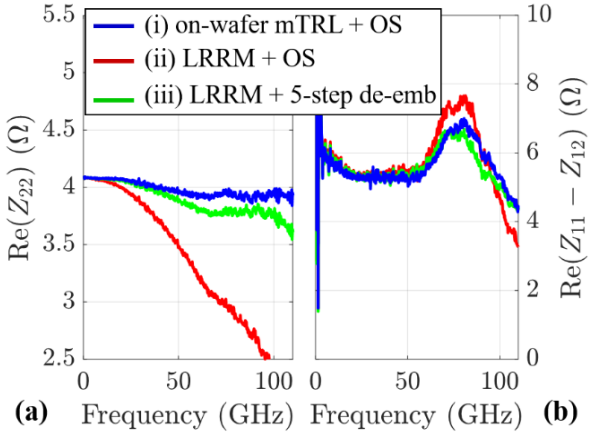

Fig. 3. $Re(Z_{22})$ (a) and $Re(Z_{11}-Z_{12})$ (b) from cold FET measurement biased in strong inversion. Comparison between different calibration/de-embedding strategies.

The classic Open-Short-M1 strategy (ii) is limited above 20 GHz, frequency at which the distributive elements of the accesses become significant and induce a decreasing $Re(Z_{22})$ trend with frequency. Although the extraction is improved with the more complex 5-step de-embedding (iii), it is still diverging above 100 GHz, whereas the 2-tier calibration followed by Open-Short de-embedding (i) displays a consistent behavior for $Re(Z_{22})$ up to 110 GHz. Whereas for the $Re(Z_{11}-Z_{12})$ curves, all measurements agree well up to 60 GHz, then they all feature a resonance in the 60-90 GHz range followed by a decreasing trend with frequency. The resonance is identified as coming from probe coupling to the on-wafer environment that is not well corrected. More details about these trends from Fig. 3(b) are provided in the next section.

V. Effect of Probe Technology

The FETs as well as all de-embedding structures have been measured with different probe technologies. This section compare the results obtained with the different probes. 3 pairs of probes with 100 µm-pitch have been used: FormFactor Infinity probes for DC-67 GHz band (Inf67), FormFactor Infinity for DC-110 GHz band (Inf110), Picoprobe GGB for DC-110 GHz band (GGB110). The measurements corrected by the 2-tier calibration followed by an Open-M1-Short-M1 de-embedding described previously are shown in Fig. 4.

Overall, there is little deviation ($\Delta Re(Z_{22}) < 8\%$ and $\Delta Re(Z_{11}-Z_{12}) < 20\%$ at 20 GHz) among the measurements with different probes. It means that all elements (FET, de-embedding and calibration structures) have been measured with high quality and repeatable contact and that there is little die-to-die process variation. Fig. 4(a) shows significantly decreasing $Re(Z_{22})$ curves at high frequency with the Infinity probes (both Inf67 and Inf110), while the curve features up to 5% variation from its value at 20 GHz with the GGB probes.

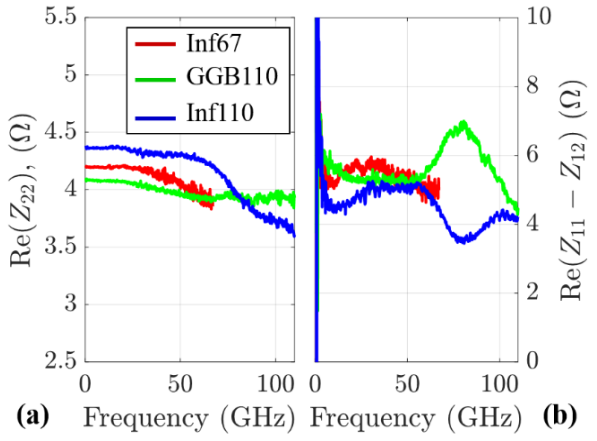

Fig. 4. $Re(Z_{22})$ (a) and $Re(Z_{11}-Z_{12})$ (b) from cold FET measurement biased in strong inversion. Comparison between measurements from different probes.

Fig. 4(b) instead presents curves with some strong resonances resulting in variations up to ~30% around its nominal value. Fig. 5(b) shows the peak f_{max} (= f.√U, U being the unilateral power gain and f the frequency) measurements (normalized to nominal f_{max} of the technology ~400 GHz), which feature the same resonances as in the $Re(Z_{11}-Z_{12})$ curves, with variations also about 25% from its nominal value. These resonances are not part of the FET behavior. Indeed, they are also measured in the on-wafer calibration structures in similar frequency ranges: see Fig. 5(a) for the Open measurements. The resonances in measurements represent energy loss in the system (that is not reflected back, nor transmitted to the other port). Their intensity and frequency vary according to the probe technology. This energy loss is therefore attributed to probe coupling with the nearby on-wafer environment. Indeed measurements of similar Open-M1 de-embedding structures at different positions on the chip, i.e. with different neighboring structures, entail resonances at different frequencies and magnitude, as shown in Fig. 6.

From Fig. 5, we see that Infinity probes coupling with neighboring structures causes a resonance at relatively low

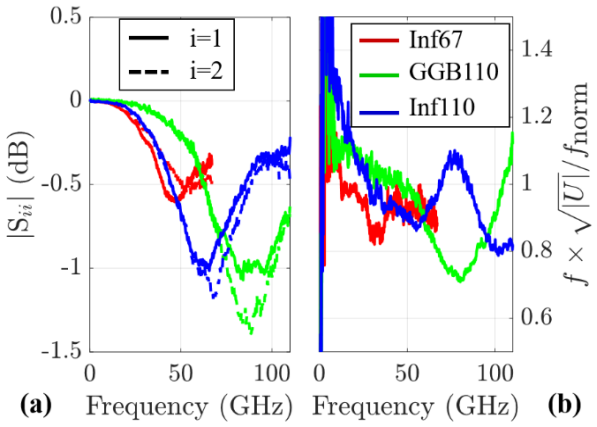

Fig. 5. (a) Return loss from on-wafer calkit Open measurements. (b) f_{max} versus frequency of the FET normalized to the nominal f_{max} of the technology. Comparison between measurements from different probes.

frequencies and significant deviation in measurements occur already starting from 20-30 GHz. Whereas, although the resonance is stronger with the GGB probes, the measure-

ments show small variation below 50-60 GHz. These probes "signatures" have recently been identified for the Inf110 and GGB110 probes in [2]. Indeed, the Inf110 features a solder joint about 500 µm away from the probe tips in the horizontal direction that couples with the underneath ground plane on the chip. The GGB110 probe instead couples with the chip ground plane along the CPW line-like tips in a more distributed way. The strong and more localized resonance present in the Inf110 measurements (observed in Figs. 5 and 6 for instance) is explained by the probe localized coupling with chip ground plane beneath the soldering point, while the more distributed resonance at higher frequencies with the GGB probe is explained by a more distributed probe coupling [2].

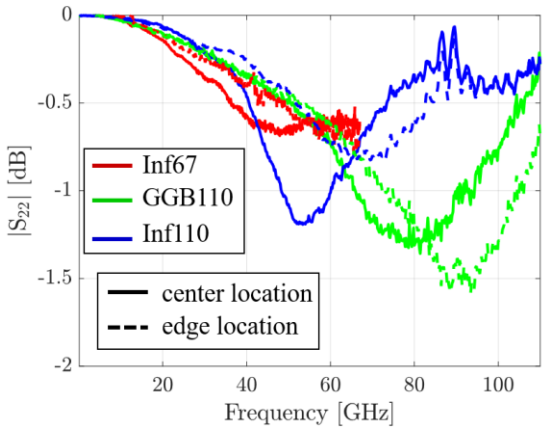

Fig. 6. $|S_{22}|$ of Open-M1 de-embedding structure measurement with different probes on 2 different locations on the chip: in the center and on one edge.

The steady frequency drop in $Re(Z_{22})$ (cf. Fig. 4(a)) for the Inf110 probes above 70 GHz and in $Re(Z_{11}-Z_{12})$ (cf. Fig. 4(b)) for the GGB110 probes above 90 GHz could be related to some distributed probe to probe crosstalk not (well) accounted for in the calibration and de-embedding procedure [2].

Despite the effort to reproduce a similar neighboring environment to each test structure (placing them in an array, such that each structure sees another one on the left and right at a constant distance, adding a ground plane covering the whole chip) and to avoid slot modes between the grounds of adjacent structures, we are still subject to probe coupling with on-wafer environment that induces resonances thereby deteriorating measurements. Since the probe coupling to the neighboring environment differs from one structure to another according to its position, the corrected measurements still suffer from some resonances. For the Inf67 probes, the coupling is sufficiently weak to be well corrected by the on-wafer calibration process, as seen in Fig. 4(b). The GGB110 probes entail a strong coupling at higher frequencies, but yield nonetheless consistent measurements up to 50-60 GHz, resulting in clean measurement correction (< ±3% variation from 10 to 60 GHz in $Re(Z_{11}-Z_{12})$ curve) up to these frequencies. The strong coupling at higher frequencies (> 60 GHz) could not be entirely corrected and is still present in the de-embedded measurements. The Inf110 probes also feature a strong coupling, although not as strong as the GGB110 probes. However, the Inf110 coupling affects measurements at lower frequencies (20-30 GHz) and over a wider frequency range (~20-90 GHz) than the GGB110 probes. As a result, the corrected measurements with Inf110 probes feature small variations < ±19% in $Re(Z_{11}-Z_{12})$ from ~20-90 GHz (and < ±8% from 10 to 60 GHz) due to probe coupling as can be seen in Fig. 4(b).

Having a ground plane covering the whole chip has been proposed in [7] to reduce probe coupling with the substrate and slot mode excitation. Despite increasing the probe coupling with the chip, the idea of presenting an infinite ground plane to the probes is to ensure an identical probe coupling for each test structure such that it could be theoretically corrected with the on-wafer calibration. However, the chip dimension is limited in practice and so is the ground plane. Due to this limitation, the probe coupling with the neighbor environment changes according to the test structure position on the chip and can therefore not be entirely removed from measurements. For those reasons, a practical way to reduce probe coupling and its detrimental effect on measurements is to avoid designing a ground plane covering the chip and to increase spacing between test structures.

VI. Conclusion

Extracting transistor parameters such as the extrinsic gate resistance is not straightforward and at high frequency (f > ~20 GHz) is very sensitive to the choice of calibration and de-embedding procedure as well as to the choice of probe techology used for measurements, with a strong correlation to the f_{max} curve. We have confirmed here that the most accurate extraction is achieved with the 2-tier calibration followed by a classic Open-Short de-embedding. The choice of probe technology is also important, as the probe coupling with the underneath ground plane on the chip is responsible for strong resonances in measurements that cannot be fully corrected by the on-wafer calibration and still persist in the extracted series resistances. The best measurements are realized with DC-110 GHz Picoprobe probes with clean extraction up to ~60 GHz (< ±3% variation in $Re(Z_{11}-Z_{12})$ from 10 to 60 GHz). The measurements with DC-110 GHz Infinity probes fluctuate significantly starting from 20 GHz, while the DC-67 GHz Infinity probes measurements deviate less and are the least affected from probe coupling with on-wafer environment.

Acknowledgment

The authors thank GlobalFoundries for chip fabrication and support. This work was partially supported via Beyond5 ECSEL project. Lucas Nyssens is a research fellow of the Fonds de la Recherche Scientifique (FNRS).

References

[1] B. Saha *et al.*, *IEEE EDL*, vol. 42, no. 1, pp. 14-17, Jan. 2021.

[2] S. Fregonese *et al.*, *IEEE TED*, vol. 68, no. 12, pp. 6007-6014, Dec. 2021.

[3] J. C. Tinoco and J.-P. Raskin, *2008 ICCDCS*, 2008, pp. 1-6.

[4] J. Rimmelspacher *et al. 2019 ARFTG*, 2019, pp. 1-4.

[5] R. B. Marks, *IEEE T-MTT*, vol. 39, no. 7, pp. 1205-1215, July 1991.

[6] C. Raya, "*Modélisation et optimisation de transistors bipolaires à hétérojonction Si/SiGeC ultra rapides pour applications millimétriques*," PhD thesis, University of Bordeaux, Bordeaux, 2008.

[7] M. Cabbia *et al.*, *IEEE TED*, vol. 67, no. 12, pp. 5639-5645, Dec. 2020.

[8] D. F. Williams *et al. IEEE Trans. on Terahertz Sci. and Technol.*, vol. 3, no. 4, pp. 433-439, July 2013.

[9] J. C. Tinoco and J.-P. Raskin, *2008 EuMIC*, 2008, pp. 127-130.

[10] D. F. Williams *et al.*, *IEEE T-MTT*, vol. 62, no. 3, pp. 658-668, 2014.

[11] N. Derrier *et al.*, *2012 IEEE BCTM*, 2012, pp. 1-8.